**Kauno technologijos universitetas**

Matematikos ir gamtos mokslų fakultetas

# Lietuvos būsto kainų modeliavimas panaudojant didžiųjų duomenų analitikos priemones

Baigiamasis magistro studijų projektas

**Eidisonas Gridziuška**

Projekto autorius

**Doc. dr. Arvydas Jadevičius**

Vadovas

**Doc. dr. Mindaugas Kavaliauskas**

Vadovas

**Kaunas, 2025**

**Kauno technologijos universitetas**

Matematikos ir gamtos mokslų fakultetas

# Lietuvos būsto kainų modeliavimas panaudojant didžiųjų duomenų analitikos priemones

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

**Eidisonas Gridziuška**

Projekto autorius

**Doc. dr. Arvydas Jadevičius**
Vadovas

**Doc. dr. Mindaugas Kavaliauskas**
Vadovas

**Doc. dr. Andrius Grybauskas**
Recenzentas

**Dr. Mindaugas Bražėnas**
Recenzentas

**Kaunas, 2025**

**Kauno technologijos universitetas**

Matematikos ir gamtos mokslų fakultetas

Eidisonas Gridziuška

# Lietuvos būsto kainų modeliavimas panaudojant didžiųjų duomenų analitikos priemones

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Eidisonas Gridziuška

*Patvirtinta elektroniniu būdu*

Gridziuška Eidisonas. Lietuvos būsto kainų modeliavimas panaudojant didžiųjų duomenų analitikos priemones. Magistro studijų baigiamasis projektas / vadovas doc. dr. Arvydas Jadevičius, vadovas doc. dr. Mindaugas Kavaliauskas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai).

Reikšminiai žodžiai: būsto kainos, aruodas.lt, daugialypė regresija, mašininis mokymasis, sprendimų medžiai.

Kaunas, 2025. 72 p.

## Santrauka

Lietuvos būsto kainų dinamika yra itin svarbi tiek ekonominiu, tiek socialiniu požiūriu, nuolat kylančios būsto kainos ne tik mažina gyventojų perkamąją galią, bet ir lemia tai, kad būsto išlaidos užima vis didesnę namų ūkio biudžeto dalį. Tokios rinkos sąlygos verčia gyventojus atsakingai vertinti sprendimus dėl būsto pasirinkimo, nes dažnai šie sprendimai yra susiję su svarbiausiais finansiniais įsipareigojimais. Neįvertinus visų alternatyvų ir jų potencialo, galima susidurti su rimtomis pasekmėmis – finansiniu nestabilumu ar net neigiamu turtiniu balansu. Dėl šių priežasčių vis svarbesnis tampa tikslus, duomenimis paremtas būsto vertės nustatymas. Nors šiame darbe nagrinėjami ir būsto kainų indeksai bei jų pokyčius lemiantys makroekonominiai veiksniai, pagrindinis dėmesys skiriamas individualių NT objektų kainos modeliavimui. Tyrimas apėmė butų duomenų rinkimą iš „aruodas.lt“ skelbimų portalo, taip pat žvalgomąją analizę, požymių inžineriją, optimalių struktūrinių parametrų paiešką bei regresinių modelių kūrimą ir jų hiperparametrų optimizavimą. Įvertinus modelių tikslumą pagal *RMSE* ir *MAPE* rodiklius, nustatyta, kad geriausi rezultatai visuose duomenų rinkiniuose pasiekti taikant sprendimų medžių ansamblių algoritmus. Geriausi modeliai panaudoti atliekant paaiškinamųjų požymių svarbumo analizę. Naudojant butų pardavimo duomenis, nustatyta, kad svarbiausi kvadratinio metro kainos paaiškinamieji požymiai yra pastato amžius ir Euklido atstumas iki miesto centro. Tyrimo pabaigoje atliktas butų kainų modeliavimas, naudojant skirtingus, pagal svarbiausius paaiškinamuosius požymus gautus segmentus. Tai leido įvertinti modelių tikslumą skirtingose rinkos nišose ir pateikti praktines rekomendacijas gyventojams, planuojantiems įsigyti butą. Be to, tyrime modeliuotos ir nuomos kainos, o tai sudarė galimybę parduodamiems butams apskaičiuoti preliminarius pardavimo ir nuomos kainų santykius. Šie rezultatai padėjo įsivertinti, per kiek laiko investicijos į konkretaus segmento butą atsipirktų vien iš nuomos generuojamų pajamų, tokiu būdu gyventojams suteikta papildoma vertinimo perspektyva, įtraukiant ir investicinį aspektą.

Gridziuška Eidisonas. Lithuanian Housing Prices Modelling with the Use of Big Data Analytics Tools. Master's Final Degree Project / supervisor doc. dr. Arvydas Jadevičius, supervisor doc. dr. Mindaugas Kavaliauskas; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences).

Keywords: housing prices, aruodas.lt, multiple regression, machine learning, decision trees.

Kaunas, 2025. 72 pages.

## Summary

The dynamics of housing prices in Lithuania are highly important from both economic and social perspectives. Constantly rising prices not only erode household purchasing power but also result in housing expenses taking up an increasingly larger share of household budgets. Such market conditions compel individuals to carefully evaluate their housing choices, as these decisions are often tied to their most significant financial commitments. Failure to assess all alternatives and their potential can lead to serious consequences, including financial instability or even negative equity. Therefore, accurate, data-driven housing valuation is becoming increasingly important. While this study includes an overview of housing price indices and the macroeconomic factors influencing their trends, the primary focus is placed on modelling the value of individual real estate properties. The research involved collecting data from the real estate portal aruodas.lt, conducting exploratory data analysis, performing feature engineering, searching for optimal structural parameters, and developing regression models with tuned hyperparameters. Model performance was evaluated using RMSE and MAPE metrics, and the best results across all datasets were achieved using ensemble tree-based algorithms. The best-performing models were then used to conduct feature importance analysis. Using apartment sales data, it was determined that the most important explanatory variables for apartment square meter price were the building’s age and the Euclidean distance to the city center. Towards the end of the study, apartment price modelling was performed using segments derived from the most important explanatory variables. This allowed for the evaluation of model accuracy across different market niches and the development of practical recommendations for individuals planning to purchase an apartment. Additionally, rental prices were also modelled, enabling the calculation of approximate price-to-rent ratios for listed properties. These results provided insights into how long an investment in a particular apartment segment would take to pay off solely through rental income, thereby offering buyers an additional evaluation perspective that integrates the investment aspect of homeownership.

**Turinys**

## Lentelių sąrašas

## Paveikslų sąrašas

## Santrumpų ir terminų sąrašas

**Santrumpos:**

BKI - būsto kainų indeksas.

NT – nekilnojamas turtas.

VECM – vektorinis paklaidų korekcijos modelis (angl. *Vector Error Correction Model*).

ARDL – autoregresinis išskirstytų vėlavimų modelis (angl. *Autoregressive Distributed Lag Model*).

SVR – atraminių vektorių regresija (angl. *Support Vector Regression*).

RMSE – vidutinės kvadratinės paklaidos šaknis (angl. *Root Mean Square Error*).

MAPE – vidutinė absoliuti procentinė paklaida (angl. *Mean Absolute Percentage Error*).

OLS – mažiausių kvadratų metodas (angl. *Ordinary Least Squares Method*).

ESF – tikrinių vektorių erdvinis filtravimas (angl. *Eigenvector Spatial Filtering*).

HPM – hedonistinis kainodaros metodas (angl. *Hedonic Pricing Method*).

PSCM – pseudo dvynio palyginimo metodas (angl. *Pseudo Self Comparison Method*).

RF – atsitiktinis miškas (angl. *Random Forest*).

ANN – dirbtinis neuroninis tinklas (angl. *Artificial Neural Network*).

MLP – daugiasluoksnis perceptronas (angl. *Multilayer Perceptron*).

PCA – pagrindinių komponenčių analizė (angl. *Principal Component Analysis*).

KDE – branduolio tankio įvertinimas (angl. *Kernel Density Estimation*).

# Įvadas

Tikriausiai kiekvienas iš mūsų bent kartą gyvenime svarstėme arba svarstysime apie būstą, kaip asmeninę investiciją į komfortą, saugumą ir ateitį. Tradiciškai būstas suvokiamas kaip ilgalaikė gyvenamoji erdvė, o pirkėjui šis turtas yra susijęs su emociniu ir finansiniu stabilumu. Tačiau prieš priimant šį sprendimą, pirkėjui reikėtų atkreipti dėmesį į tai, kad pastaraisiais metais Lietuvos nekilnojamojo turto rinkoje pastebimi reikšmingi pokyčiai – augančios būsto kainos šalies didmiesčiuose, didelis skaičius naujų projektų ir padidėjusi paklausa tiek vietinių gyventojų, tiek užsienio investuotojų atžvilgiu. Tokie NT rinkos veiksniai po mažu keičia ir pačią būsto įsigijimo sampratą. Šių dienų visuomenėje, kai vis didesnė dalis gyventojų naudojasi bankų išduodamomis paskolomis, ilgalaikėje perspektyvoje į būstą žvelgiama ne tik kaip į gyvenamąją erdvę, bet ir kaip į finansinį turtą, kuris gali generuoti nuomos pajamas. Dėl to, dinamiškoje NT rinkos aplinkoje svarbu gebėti prognozuoti būsto kainas ir suprasti pagrindinius kainų elgsenai įtakos turinčius veiksnius. Tam dažnai pasitelkiami būsto kainų indeksai su atitinkamais laiko eilučių analizės metodais, visa tai leidžia nagrinėti bendras rinkos tendencijas ir ilgalaikes kryptis. Vis dėlto, vertinant konkretų objektą, šie metodai dažnu atveju yra per daug apibendrinti, todėl neatspindi kiekvieno būsto ir jo ypatybių. Dėl šios priežasties daugiau dėmesio tiek akademinėje, tiek praktinėje veikloje skiriama rinkos palyginimo metodams, kurie leidžia į būsto kainą žvelgti per požymių visumos prizmę. Mokslinėje literatūroje daug dėmesio sulaukia hedonistine vertinimo kainodara paremti regresiniai modeliai, kurie leidžia prognozuoti būsto kainą, atsižvelgiant į jo struktūrinius, lokacinius ar aplinkos požymius. Tokie metodai leidžia ne tik priimti racionalius, duomenimis pagrįstus sprendimus, bet ir identifikuoti svarbiausius veiksnius, formuojančius būsto kainą. Taip pat svarbu paminėti, kad kartu su augančiu susidomėjimu pažangesniais prognozavimo metodais, sparčiai didėja ir prieinamos informacijos kiekis bei požymių įvairovė – remiantis šių laikų technologijomis, į modeliavimo procesą gali būti įtraukiama tiek skelbimų aprašymuose esanti tekstinė, tiek nuotraukose esanti vaizdinė informacija. Visa tai sudaro palankias sąlygas didžiųjų duomenų analitikos priemonių taikymui. Šios priemonės naudojamos beveik visuose būsto kainos modeliavimo etapuose – nuo žvalgomosios analizės ir požymių inžinerijos iki pažangių mašininio mokymosi modelių kūrimo.

**Darbo problema**: daugeliui NT rinkos dalyvių, ypač mažmeniniams investuotojams ar pirmąjį būstą perkantiems gyventojams, trūksta priemonių, leidžiančių objektyviai įvertinti, ar konkretaus būsto kaina yra racionali. Atsižvelgiant į tai, kad būsto įsigijimas dažnu atveju yra susijęs su ilgalaikiais finansiniais įsipareigojimais, galima teigti, kad tai yra vienas svarbiausių finansinių sprendimų pirkėjo gyvenime. Šis sprendimas tampa dar sudėtingesnis, kai įtraukiamas ir investicinis aspektas – tokiais atvejais svarbi ne tik būsto pardavimo kaina, bet ir likvidumas bei nuomos gereneruojamos pajamos. Neturint patikimo, duomenimis pagrįsto rinkos vertinimo įrankio yra labai sudėtinga pasirinkti optimalų būstą tiek gyvenamuoju, tiek investiciniu požiūriu.

**Darbo tikslas**: kurti didžiaisiais duomenimis grįstus regresinius modelius, leidžiančius prognozuoti Lietuvos miestų butų pardavimo ir nuomos kainą.

**Darbo objektas**: viešai prieinami Vilniaus ir Kauno miestų butų pardavimo ir nuomos skelbimai iš „aruodas.lt“ svetainės.

**Darbo uždaviniai**:

1. Išanalizuoti mokslinę literatūrą, susijusią su Lietuvoje bei užsienyje atliktais, panašios tematikos tyrimais, įvardinti tyrimuose taikytus metodus, įvertinti tyrimų rezultatus ir autorių išvadas;
2. Naudojant programinę įrangą, surinkti Vilniuje ir Kaune parduodamų bei nuomojamų butų duomenis;
3. Naudojant didžiųjų duomenų analitikos priemones, atlikti surinktų duomenų žvalgomąją analizę;
4. Naudojant skirtingus požymių rinkinius, sukurti ir ištestuoti regresinius modelius, parinkti optimalius modelių hiperparametrus;
5. Atrinkti didžiausiu tikslumu pasižyminčius modelius, naudojant juos, identifikuoti svarbiausius paaiškinamuosius požymius;
6. Suformuluoti išvadas ir pateikti praktines rekomendacijas.

# 1. Literatūros apžvalga

## 1.1. Būsto kainų reikšmė, poveikis ekonomikai bei visuomenei

Išsivysčiusioje visuomenėje nekilnojamas turtas atlieka dvejopą funkciją – viena vertus, jis tenkina esminius žmogaus poreikius, kita vertus, veikia kaip kapitalo investicija. P. Tostevin ir kt. (2022) [1] ataskaitos rezultatai atskleidė, kad pasaulinė nekilnojamojo turto rinka 2022 metų pabaigoje buvo verta 379,7 trilijono JAV dolerių – ši suma beveik keturis kartus viršijo tų pačių metų globalaus bendrojo vidaus produkto (BVP) dydį (100,6 trilijono JAV dolerių). Nepaisant to, kad 2022 metais NT rinkos vertė sumažėjo 2,8% lyginant su 2021 metais, ilgesniame laikotarpyje pastebėta aiški augimo tendencija – nuo 2019 metų bendra vertė padidėjo 18,7%. Ataskaitoje taip pat pažymėta, kad gyvenamosios paskirties objektai 2022 metais sudarė net 76% visos NT rinkos vertės, ši rinkos dalis piniginiu požiūriu buvo kelis kartus didesnė lyginant su kitoms turto klasėmis, pavyzdžiui, akcijomis, obligacijomis, o P. Tostevin ir kt. teigimu, ateityje šis skirtumas tik didės [1]. Visa tai patvirtina, kad nekilnojamas turtas, o ypač gyvenamosios paskirties objektai, išlieka kaip dominuojantis finansinis instrumentas. Tačiau reikia atkreipti dėmesį į tai, kad kiekvienas tokio tipo instrumentas turi atitinkamą riziką. Nors būsto rinka yra viena iš pagrindinių globalios ekonomikos dalių, ji taip pat yra pažeidžiama. 2007-2009 metų pasaulinė finansų krizė, prasidėjusi JAV *subprime* hipotekos rinkoje, parodė, kokius su ekonominiu nestabilumu susijusius padarinius gali sukelti reikšmingi pokyčiai būsto rinkose ir jų funkcionavimo mechanizmuose. Krizės metu buvo paveikti ne tik milijonai namų ūkių, netekusių būsto ar susidūrusių su nepakeliamomis paskolų sąlygomis, bet ir visos finansinės institucijos, kurių stabilumas tiesiogiai priklausė nuo nekilnojamojo turto vertės. Kaip pastebi D. Maclennan ir kt. (2023) [2], vienas iš ryškiausių pavyzdžių – Airijos būsto rinkos žlugimas, kai nekilnojamojo turto kainos nuo 2007 iki 2012 metų sumažėjo daugiau nei 50%, o bankų patirti nuostoliai tapo tokie dideli, kad valstybė buvo priversta įsikišti ir suteikti finansinę paramą, kurios dydis prilygo net 40% šalies bendrojo vidaus produkto. Autoriai taip pat akcentuoja, kad iš 50 sisteminių bankų krizių per pastaruosius dešimtmečius, daugiau nei puse buvo susijusios su būsto kainų „burbulų“ sprogimais [2]. Tokie pavyzdžiai leidžia susidaryti nuomonę, kad nestabilios būsto rinkos padariniai apima visų išsivysčiusių valstybių finansinį saugumą bei ekonominę raidą. Dėl šios priežasties būsto kainos yra atidžiai analizuojamos bei įvardinamos kaip vienas iš svarbiausių ekonominės būklės indikatorių.

Ne mažiau svarbus yra ir socialinis būsto kainų stabilumo aspektas, ypač žvelgiant iš individualių namų ūkių perspektyvos. Šių laikų visuomenėje būstas dažnai yra ne tik pagrindinis žmogaus turtas, bet ir didžiausias ilgalaikis finansinis įsipareigojimas. Dažnai būstas įsigyjamas pasitelkiant banko paskolą, o reguliarios būsto paskolos įmokos tampa reikšminga dalimi namų ūkio biudžeto. Dėl šios priežasties būsto kainų svyravimai tiesiogiai veikia gyventojų finansinį saugumą, jų galimybes kaupti turtą ar taupyti ateičiai. Esant staigiam NT kainų nuosmukiui, gyventojai, pirkę būstą kainų piko metu, gali atsidurti neigiamos nuosavybės (angl. *negative equity*) situacijoje – kai turto vertė tampa ženkliai mažesnė už negrąžintos paskolos sumą. Tokia padėtis ne tik apsunkina finansinius sprendimus (pvz., keisti arba parduoti būstą), bet ir didina socialinę įtampą, ypač tarp jaunų šeimų ar mažas pajamas turinčių gyventojų, kurių galimybės prisitaikyti prie ekonominių pokyčių yra ribotos. Be to, nestabilios būsto kainos gali turėti reikšmingą poveikį darbo rinkai. M. R. Farzanegan ir kt. (2024) [3] savo straipsnyje teigia, kad augančios būsto kainos gali neigiamai paveikti OECD (angl. *Organisation for Economic Co-operation and Development*) šalių darbuotojų produktyvumą. Autoriai išskiria tris pagrindines situacijas, kuriose būsto kainos veikia darbo našumą [3]:

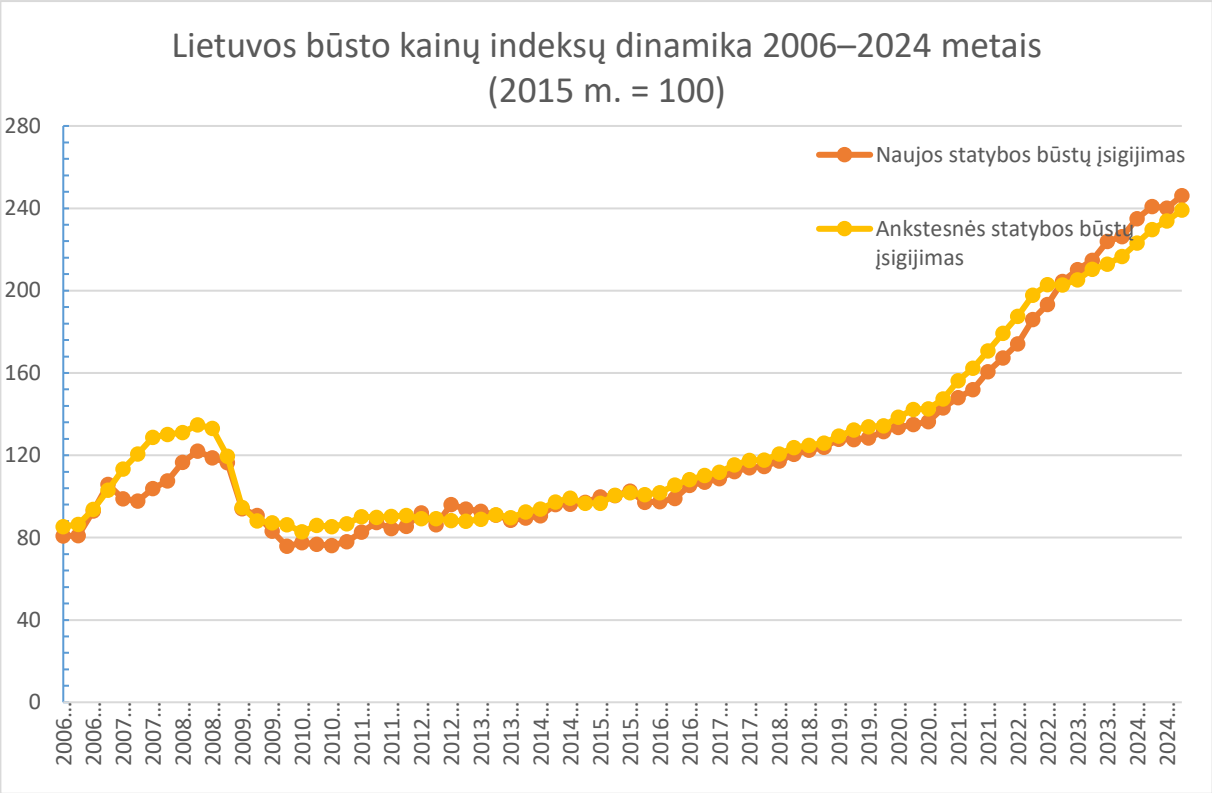
- žmogiškojo kapitalo formavimas – didėjančios būsto išlaidos gali nukreipti namų ūkių lėšas nuo investicijų į išsilavinimą, kuris yra vienas iš pagrindinių darbo produktyvumo veiksnių;
- darbuotojų mobilumas – didėjant būsto kainoms ir nuomos mokesčiams, žemesnes pajamas gaunantys gyventojai yra priversti persikelti toliau nuo darbo vietų, tokiu būdu mažėja darbo efektyvumas ir produktyvumas;
- kapitalo paskirstymas – didėjant būsto kainoms, vis didesnė dalis gyventojų investicijų yra nukreipiama į nekilnojamąjį turtą, tokiu būdu mažėja kapitalo prieinamumas inovatyvesniems valstybės sektoriams.

Literatūroje taip pat analizuojama ir kita, itin aktuali problema – mažėjantis būsto įperkamumas. Kaip pažymi D. Schwartz (2016) [4], pastaraisiais metais daugelio šalių bendruomenės susiduria su sparčiai augančiomis būsto kainomis, kurios didėja greičiau nei gyventojų pajamos. Dėl šio disbalanso vis daugiau namų ūkių, ypač gaunančių vidutines ar mažesnes pajamas, tampa finansiškai pažeidžiami. Remiantis D. Schwartz atlikta analize, 2014 metais JAV būsto kainos buvo net 60% didesnės nei 2000-aisiais, tuo tarpu vartotojų kainų indeksas per tą patį laikotarpį pakilo tik 37%, o vidutinės gyventojų pajamos – vos 28% [4]. Šie rezultatai aiškiai rodo, kad gyventojų perkamoji galia mažėja, o tai ilgainiui virsta sistemine problema.

Nepaisant visų minėtų problemų ir rizikų, būstas išlieka viena patraukliausių investicinių priemonių tiek privačių asmenų, tiek finansinių institucijų akyse. Lyginant su kitomis turto klasėmis – akcijomis, obligacijomis ar alternatyviomis investicijomis, nekilnojamas turtas išsiskiria tuo, kad sudaro galimybes generuoti nuolatines grynųjų pinigų srauto pajamas nuomos būdu bei išlaikyti arba didinti kapitalo vertę ilguoju laikotarpiu. Be to, į būstą dažniausiai žvelgiama kaip į stabilesnę investiciją, kuri, skirtingai nei kitos alternatyvos, yra susijusi su realiu fiziniu turtu. Šių laikų visuomenėje aiškiai pastebima ir urbanizacijos įtaka – daugelio šalių didmiesčiuose būsto paklausa sparčiai didėja, o tai reiškia, kad taip pat didėja investicijų kiekis į urbazinuotų zonų gyvenamąjį NT. Dėl šios priežasties didmiesčių būsto segmentas išsiskiria tiek savo investicine grąža, tiek socialine reikšme – sprendžiamos ir urbanistinės plėtros, gyventojų mobilumo bei socialinės infrastruktūros prieinamumo problemos. Atsižvelgiant į visas išvardintas tendencijas, galima teigti, kad stebėti ir vertinti šią rinką privalo visi – tiek politinių sprendimų priėmėjai, tiek paprasti gyventojai. Viena iš pagrindinių priemonių, leidžiančių įvertinti esamą šios rinkos būklę ir numatyti jos raidos kryptis – būsto kainų indeksai. Tolimesniame skyriuje bus nagrinėjama, kaip interpretuoti šiuos indeksus, bei kokie metodai taikomi siekiant tiksliai prognozuoti jų pokyčius.

### 1.2. Būsto kainų indekso prognozavimas

BKI leidžia įvertinti, kaip per tam tikrą laikotarpį pasikeitė bendra būsto kaina, rodiklis yra labai aktualus ekonomistams, politikos formuotojams, investuotojams, taip pat ir gyventojams, planuojantiems pirkti arba parduoti būstą. Iš esmės BKI parodo vidutinę būsto vertę laike, lyginant su nustatytais baziniais metais (pvz., 2015 m. = 100). Europoje pastaraisiais metais pastebimas nuolatinis šio indekso augimas. Ne išimtis ir Lietuva – BKI augimas pastebimas tiek Vilniuje, tiek mažesniuose šalies miestuose. 1 pav. pateiktas grafikas iliustruoja Lietuvos būsto kainų indeksų pokyčius nuo 2006 iki 2024 metų. Grafike pastebimi 2008 metų finansų krizės padariniai – krizės metu indeksų reikšmės ženkliai sumažėjo. Vėliau, nuo 2013 metų, pastebimas stabilus būsto kainų augimas, kuris ypač suintensyvėjo nuo 2021 metų. Duomenys paimti iš Lietuvos oficialios statistikos portalo, skirtingi būsto kainų indeksai sudaryti atitinkamai naudojant naujos ir ankstesnės statybos būstų įsigijimo duomenis [5].




**1 pav.** Lietuvos būsto kainų indeksai 2006–2024 metais

Ekspertai išskiria keletą indekso augimo priežasčių – besikeičianti geopolitinė situacija, padidėjusi būsto paklausa, rekordinis infliacijos lygis, centrinio banko palūkanų normų kintamumas ir kt. Kaip teigia Dr. Sabyasachi Tripathi (2019) [6], būsto kainoms reikšmingą įtaką daro įvairūs makroekonominiai veiksniai, tarp jų – nuomos kainos, pinigų perkamoji galia, bendrasis vidaus produktas, infliacija, taip pat ir valiutos keitimo kursai. Tyrimo autorius pabrėžia, kad didėjantis BVP lemia būsto kainų augimą – kai ekonomika auga, gyventojų pajamos didėja, todėl jie gali skirti daugiau lėšų būstui įsigyti. Tuo tarpu infliacija veikia dviprasmiškai - viena vertus, ji mažina perkamąją galią ir gali riboti būsto įperkamumą, tačiau kita vertus – didina būsto kainas dėl augančių statybos išlaidų. Panašų poveikį daro ir pinigų pasiūla – jei centriniai bankai vykdo laisvesnę pinigų politiką ir palūkanų normos išlieka žemos, skolinimasis yra labiau prieinamas, o tai skatina būsto paklausą ir kainų augimą. Lietuvos būsto rinkos kontekste situacija iš esmės atspindi tarptautines tendencijas, tačiau čia išskiriami ir specifiniai veiksniai. L. Tupėnaitės ir kt. (2017) [7] tyrimas, nagrinėjęs būsto kainų svyravimų priežastis Lietuvoje 2005–2015 m. laikotarpiu, atskleidė, kad būsto kainų dinamikai svarbiausi yra ekonominiai veiksniai, rinkos rodikliai ir netgi kai kurie neracionalūs (psichologiniai) aspektai. Tyrimo autoriai įvardina palūkanų normas bei naujai išduodamas būsto paskolas kaip daugiausia įtakos turinčius veiksnius. Be finansinių rodiklių, būsto rinkos aktyvumą taip pat atspindi nekilnojamojo turto sandorių skaičius, kuris leidžia įvertinti pirkėjų ir pardavėjų elgseną rinkoje. L. Tupėnaitė ir kt. pažymi, kad neracionalūs veiksniai, tokie kaip vartotojų lūkesčiai ir spekuliacinė paklausa, gali prisidėti prie būsto kainų burbulo susiformavimo [7]. Kai pirkėjai tikisi, kad kainos ir toliau kils, jie yra linkę greičiau priimti sprendimus įsigyti būstą, taip dar labiau didindami paklausą ir kainų augimą. Bendru atveju būsto kainų indekso prognozavimas apibrėžiamas kaip laiko eilučių analizės uždavinys. Kadangi būsto kainos kinta laike ir priklauso nuo įvairių makroekonominių bei rinkos veiksnių, indekso prognozavimui taikomi įvairūs ekonometrijos modeliai, gebantys identifikuoti tiek ilgalaikes tendencijas, tiek trumpalaikius svyravimus. Puikus

pavyzdys – L. Naruševičiaus ir kt. (2019) [8] atliktas tyrimas, kuriame nagrinėjamos įvairios Lietuvos būsto kainų indekso modeliavimo ir prognozavimo metodikos. Autoriai taiko vektorių paklaidų koregavimo modeliavimą[1] (angl. *Vector Error Correction Modelling*), kuris leidžia nustatyti ilgalaikį ryšį tarp būsto kainų ir pagrindinių makroekonominių rodiklių, tokių kaip realios palūkanų normos, statybų sąnaudos, kreditavimo ir investicijų santykis bei namų ūkių įsiskolinimas. Pasak L. Naruševičiaus ir kt., pagrindinis *VECM* modelio pranašumas yra tas, kad jis suteikia galimybę identifikuoti laikotarpius, kai būsto kainos gali būti pervertintos ar nepakankamai įvertintos rinkos dalyvių [8]. Autorių įvardintas *VECM* modelis apima penkis kintamuosius:

- būsto kainų indeksą (*hpi*);
- realią būsto paskolų palūkanų normą (*r*);
- statybos sąnaudų kainų indeksą (*ccpi*);
- naujų būsto paskolų ir nominalių būsto investicijų santykį (*credInv*);
- esamų būsto paskolų ir nominalaus BVP santykį (*debtRatio*).

Pritaikius *VECM* modelį išsiaiškinta, kad būsto kainoms didžiausią ilgalaikį poveikį daro statybos sąnaudų kainų indeksas (*ccpi*). Modelis parodė, kad 1% padidėjęs statybos sąnaudų kainų indeksas vidutiniškai padidina būsto kainų indeksą 1,6%. Tai patvirtina hipotezę, kad statybos kaštai yra tiesiogiai perduodami į galutines būsto kainas, o tai reiškia, kad nekilnojamojo turto kainų pokyčius dažnai lemia ne tik paklausos veiksniai, bet ir pasiūlos pusėje vykstantys procesai. Be statybos sąnaudų kainų indekso modelis taip pat atskleidė reikšmingą palūkanų normų poveikį būsto kainoms – nustatyta, kad 1 procentinio punkto padidėjimas realiose būsto paskolų palūkanų normose vidutiniškai sumažina būsto kainų indeksą 1,1% [8].

Nors *VECM* modelis padeda identifikuoti ilgalaikę pusiausvyrą tarp makroekonominių rodiklių ir būsto kainų, trumpalaikių pokyčių analizei bei tikslesniam būsto kainų augimo veiksnių įvertinimui labiau tinkamas yra autoregresinis išskirstytų vėlavimų modelis (angl. *Autoregressive Distributed Lag Model*). Skirtingai nei *VECM* modelis, *ARDL* modelis nereikalauja, jog visi kintamieji būtų stacionarios laiko eilutės. Tai suteikia daugiau lankstumo empirinėje analizėje, ypač kai duomenų imtis yra ribota, kaip dažnai būna su ketvirtiniais ar mėnesiniais duomenimis. Be to, *ARDL* modelis yra naudingas formuojant trumpalaikes prognozes ir taip pat leidžia įvertinti, kurie kintamieji turi daugiausia įtakos būsto kainų augimui per artimiausius periodus. Tyrimo autoriai į *ARDL* modelį įtraukia 12 skirtingų paaiškinamųjų kintamųjų, kurie gali būti išskaidyti į tris grupes (žr. 1 lentelę).

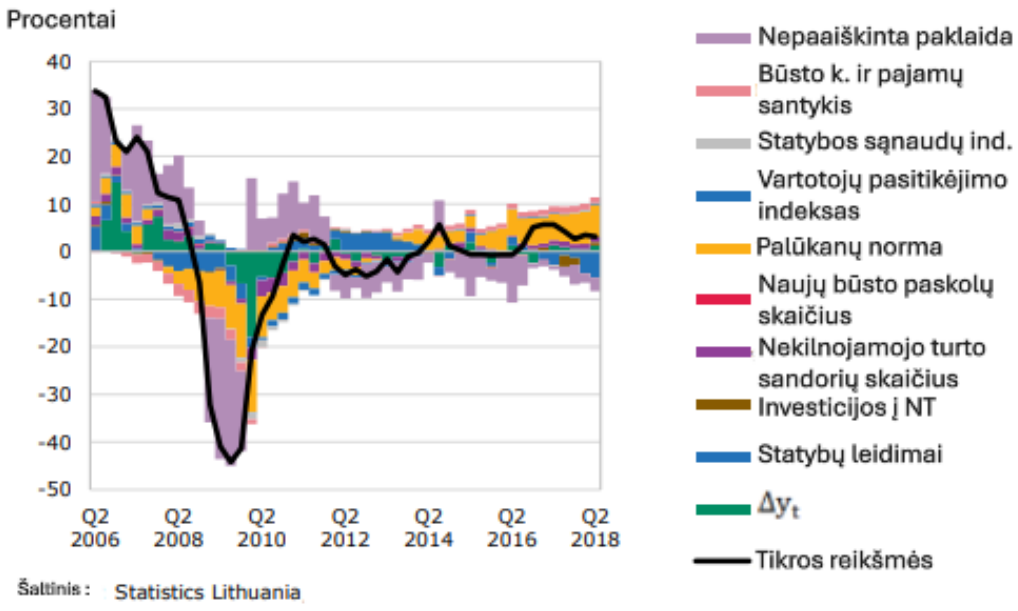
**1 lentelė.** L. Naruševičiaus ir kt. panaudoto *ARDL* modelio paaiškinamųjų kintamųjų sąrašas [8]

| **Paklausos veiksniai** | **Pasiūlos veiksniai** | **Finansiniai veiksniai** |
|---|---|---|
| Populiacijos dydis<br>Nedarbingumo lygis<br>Atlyginimas<br>Vartotojų kainų indeksas<br>Vartotojų pasitikėjimo indeksas<br>Būsto kainos ir pajamų santykis | Statybų leidimai<br>Investicijos į NT<br>Statybų sąnaudų kainų indeksas<br>Nekilnojamojo turto sandorių skaičius | Naujų būsto paskolų palūkanų norma<br>Naujų būsto paskolų skaičius |

Norėdami sumažinti galimų *ARDL* modelių skaičių ir užtikrinti teorinį nuoseklumą, tyrimo autoriai apskaičiuotiems kintamųjų koeficientams pritaikė ženklų apribojimus. Nustatyta, kad griežtai teigiamą poveikį būsto kainoms turi – gyventojų skaičius, atlyginimas, vartotojų kainų indeksas,

[1] https://en.wikipedia.org/wiki/Error_correction_model

vartotojų pasitikėjimo indeksas, statybos sąnaudų kainų indeksas ir naujų būsto paskolų skaičius. Tuo tarpu griežtai neigiamas poveikis priskirtas nedarbo lygiui, statybos leidimų skaičiui, investicijoms į NT bei naujų paskolų palūkanų normoms. Svarbu paminėti, kad NT sandorių skaičiui bei būsto kainų ir pajamų santykiui ženklų apribojimai nebuvo taikomi – tai reiškia, kad šių veiksnių įtaka būsto kainoms galėjo būti tiek teigiama, tiek neigiama. Galutinei analizei buvo atrinkti 18 individualių *ARDL* modelių. Šie modeliai vėliau buvo sujungti į vieną svorinį modelį, kuris leido įvertinti paaiškinamųjų kintamųjų įtaką būsto kainų augimo nuokrypiams nuo ilgalaikės tendencijos. Pagal 2 pav. pateiktus rezultatus galima pastebėti, kad didžiausią įtaką būsto kainų pokyčiams turėjo palūkanų normos – autoriai pabrėžia, kad būtent žemų palūkanų normų aplinka paskatino ankstesnio laikotarpio būsto kainų augimą. Iš pasiūlos veiksnių ryškiausią poveikį turėjo statybos leidimų skaičius. Tuo tarpu tokie veiksniai kaip vartotojų kainų indeksas, atlyginimai, gyventojų skaičius ir nedarbo lygis nebuvo įtraukti į galutinius modelius – tai reiškia, kad šių veiksnių įtaka būsto kainų kintamumui nebuvo statistiškai pagrįsta [8]. Įdomu ir tai, kad dalis būsto kainų pokyčių, ypač finansinės krizės laikotarpiu, liko nepaaiškinta modelio – visa tai leidžia susidaryti nuomonę, kad reikšmingos įtakos turėjo ir neekonominiai veiksniai, pavyzdžiui, vartotojų lūkesčių pasikeitimai arba psichologinė reakcija į rinkos neapibrėžtumą.



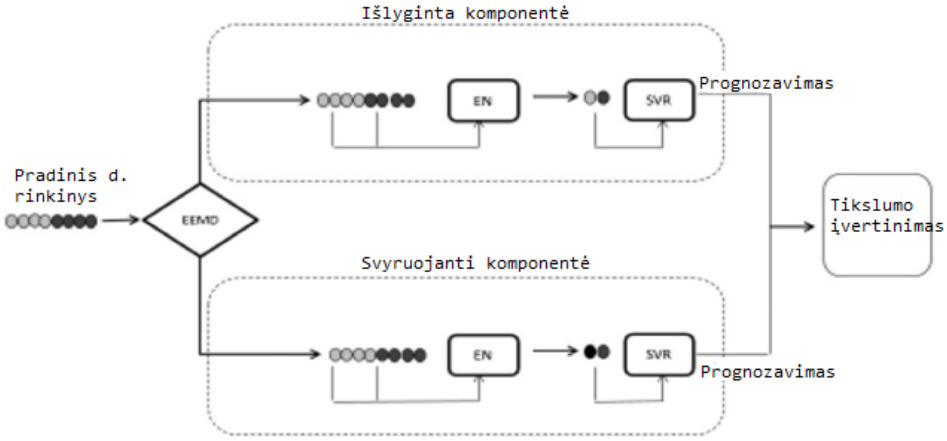

**2 pav.** Ankstesnių laikotarpių būsto kainų veiksnių vertinimas panaudojant L. Naruševičiaus ir kt. apibendrintą *ARDL* modelį [8]

Vertindami *ARDL* modelių prognozių tikslumą, autoriai taikė kelias metrikas, tokias kaip vidutinė kvadratinė paklaida (angl. *RMSE*), vidutinė absoliutinė paklaida (angl. *MAE*) bei vidutinė absoliutinė procentinė paklaida (angl. *MAPE*). Šie kriterijai leido objektyviai palyginti skirtingas modelių specifikacijas ir įvertinti modelių gebėjimą atkurti realias būsto kainų reikšmes prognozuojamu laikotarpiu. Kadangi nei vienas iš 18 išbandytų *ARDL* modelių nebuvo akivaizdžiai geriausias pagal visas vertinimo metrikas, tyrimo autoriai nusprendė taikyti prognozių kombinavimo metodą. Tai reiškia, kad vietoje vieno modelio pasirinkimo buvo naudojamas kelių geriausių modelių prognozių vidurkis. Rezultatai parodė, kad modelių kombinavimas padidino bendrą prognozavimo tikslumą [8].

Tęsdami analizę, Naruševičius ir kt. toliau plėtojo trumpalaikio ir vidutinio laikotarpio prognozavimo galimybes, taikydami vienmačius ir daugiamačius laiko eilučių modelius. Modeliavimas pradėtas nuo paprasto naivaus metodo, kuriame daroma prielaida, kad būsto kainų indekso reikšmė ateityje nesikeis – šis metodas buvo taikomas kaip lyginamasis standartas (angl. benchmark). Į vienmačių modelių rinkinį autoriai įtraukė 36 skirtingus *ARMA* modelius. Daugiamačiams modeliams autoriai pritaikė 7 skirtingus kintamųjų rinkinius, iš viso išbandyti 84 *VAR*, 84 *BVAR* bei 84 *BVAR-SV* (*BVAR* su stochastiniu kintamumu) modeliai. Vienmačių modelių atveju geriausiai pasirodė *ARMA*(4,1) modelis, o iš daugiamačių modelių geriausi buvo *VAR*(1), *BVAR*(1) ir *BVAR-SV*(1). Panašiai kaip ir ankstesniuose etapuose, autoriai taikė prognozių kombinavimą, tik šį kartą išbandytos keturios skirtingos strategijos. Pirmosios dvi – tai paprastas vidurkis ir mediana, kurie yra apskaičiuojami naudojant atskiras modelių prognozes skirtingiems horizontams. Trečioji kombinavimo strategija remiasi kiekvieno modelio aštuonių ketvirčių horizonto prognozavimo tikslumu – kuo mažesnė modelio paklaida (RMSE), tuo didesnį svorį jis gauna kombinuotoje prognozėje. Ketvirtoji strategija yra labai panaši, tačiau į kombinuotą prognozę įtraukiami tik tie modeliai, kurių santykinė vidutinė absoliutinė paklaida (angl. RMAE) yra mažesnė nei 1, t. y. jie pasirodė geriau nei lyginamasis standartas. Tyrimo rezultatai atskleidė, kad naudojant RMSE pagrįstų svorių metodiką buvo dar kartą padidintas bendras prognozių tikslumas [8].

Tęsiant apžvalgą, prasminga paminėti ir A. Jadevičiaus ir kt. (2014) [9] atliktą tyrimą, kuriame bandyta identifikuoti ciklišką Lietuvos būsto rinkoje. Šiame darbe autoriai naudojo Ober-Haus būsto kainų indeksą (OHBI), apimantį penkis didžiausius Lietuvos miestus, ir tyrė būsto kainų pokyčius nuo 1994 iki 2013 metų. Tyrimas paremtas Hodricko-Prescotto filtro taikymu – šis metodas leido atskirti trumpalaikes ciklines indeksų komponentes nuo ilgalaikių augimo tendencijų. Atliktos analizės rezultatai parodė, kad 2014 metais Lietuvos būsto kainos pasižymėjo pakankamai aiškiu ciklišku – vidutinė trukmė tarp kainų pakilimų ir nuosmukių buvo apie 3–4 metai. Taip pat nustatyta, kad laikui bėgant keitėsi šio ciklo intensyvumas, dažnis ir trukmė, pavyzdžiui, pirmaisiais nepriklausomybės metais svyravimai buvo trumpesni ir mažesnio masto, tačiau vėliau ciklai ilgėjo ir tapo aiškiau pastebimi [9]. Šiems reiškiniams įtakos turėjo tiek lokalūs ekonominiai pokyčiai, tiek tarptautinės krizės.

Siekiant įvertinti alternatyvias prognozavimo metodikas ir jų pranašumus būsto kainų modeliavimo srityje, verta panagrinėti ir užsienyje atliktus tyrimus. Vienas tokių pavyzdžių – V. Plakandaras ir kt. (2015) [10] atliktas tyrimas, kuriame prognozuojamas JAV būsto kainų indeksas, taikant ne tik tradicinius ekonometrijos modelius, bet ir pažangius mašininio mokymosi algoritmus. Tyrime siūlytas inovatyvus požiūris į būsto kainų indekso prognozavimą, jį vertinant kaip potencialų įrankį ankstyvam perspėjimui apie galimas ekonomines krizes. Tyrime naudoti 11 JAV makroekonominių rodiklių metiniai duomenys, apimantys laikotarpį nuo 1890 iki 2012 metų. V. Plakandaras ir kt. pateikė unikalią prognozavimo metodiką, kuri sujungė empirinę ansamblinę modų dekompoziciją (angl. *Ensemble Empirical Mode Decomposition*) – metodą iš signalų apdorojimo srities su atramos vektorių regresija (angl. *Support Vector Regression*) iš mašininio mokymosi. Panašiai kaip ir A. Jadevičiaus ir kt. [9] tyrime, pirmiausia visos turimos laiko eilutės buvo išskaidytos į dvi dalis – žemo dažnio išlygintas (angl. *Smoothed*), atspindinčias ilgalaikę tendenciją, ir aukšto dažnio svyruojančias (angl. *Fluctuating*), kurios fiksavo trumpalaikio svyravimo charakteristikas (žr. 3 pav.). Kintamųjų atranka atlikta naudojant *Elastic Net* metodą, tuomet galutiniai kintamųjų rinkiniai buvo perduoti į *SVR* modelį. Siekdami išvengti modelių persimokymo, autoriai taip pat taikė 4 kartų kryžminę patikrą (angl. *4-fold cross-validation*).




**3 pav.** V. Plakandaras ir kt. pasiūlyto EEMD modelio struktūra [10]

Vertinant tyrimo rezultatus, autorių siūloma *EEMD* metodika buvo lyginama su atsitiktinio vaikščiojimo (angl. *Random Walk*) modeliu, kuris, kaip įprasta tokio tipo tyrimuose, pasitarnavo kaip lyginamasis standartas. Į palyginimo procedūrą taip pat įtraukti anksčiau minėti *BAR* ir *BVAR* tipo modeliai. Rezultatai parodė, kad nors *BAR* ir *BVAR* modeliai pasižymėjo gebėjimu analizuoti kelių laiko eilučių tarpusavio priklausomybes, *EEMD* metodika, kombinuota su *SVR* algoritmu, pasižymėjo didesniu tikslumu [10]. Tarp visų vertintų *EEMD* modelių ypač išsiskyrė *EEMD-AR-SVR*, kuriame naudotos tik būsto kainų indekso praeities reikšmės. Pagal *MAPE* kriterijų, ši specifikacija pasirodė esanti pati efektyviausia, kai prognozavimo langas yra 1 periodas, todėl vėliau buvo pritaikyta ir ilgesniam laikotarpiui prognozuoti – iki 10 periodų (šiuo atveju 10 metų).

Akivaizdu, kad būsto kainų indeksas yra itin vertingas rodiklis, leidžiantis stebėti bendras rinkos tendencijas ir įvertinti, ar tam tikras laikotarpis yra palankus būsto įsigijimui. Laiko eilučių analizė, kaip rodo [8], [9], [10] tyrimų rezultatai, padeda atskleisti cikliškus būsto kainų svyravimus ir identifikuoti galimus kainų „burbulų“ sprogimus. Vis dėlto, pastarąjį dešimtmetį visų išsivysčiusių valstybių būsto rinkos pasižymėjo beveik nenustojančiu kainų augimu, taip pat vis sunkiau pastebimas minėtas NT rinkos ciklišumas. Dėl šios priežasties vien indeksais grįsti sprendimai yra riboti – indeksai leidžia įvertinti bendra rinkos kryptį, tačiau neatsako į pirkėjui taip pat svarbų klausimą: kokį konkretų objektą verta įsigyti? Atsakant į šį klausimą atsiranda poreikis metodams, kurie leistų ne tik stebėti makrolygio tendencijas, bet ir tiksliai įvertinti individualaus būsto vertę, atsižvelgiant į jo unikalias charakteristikas – fizines, lokacines ir infrastruktūrines savybes. Šiam uždaviniui spręsti plačiai taikomi hedonistiniai vertinimo modeliai (angl. *Hedonic Pricing Model*), kurie leidžia susieti būsto vertę su jį apibūdinančiais požymiais. Pereinant nuo bendro lygmens analizės prie individualių objektų vertinimo, toliau šiame darbe bus nagrinėjama literatūra, kurioje būsto vertei nustatyti taikomi regresiniai metodai – tiek klasikiniai (pvz., daugialypė tiesinė regresija), tiek pažangesni, paremti mašininio mokymosi algoritmais.

### 1.3. Klasikiniai būsto vertės nustatymo metodai

Literatūroje galima rasti gausybę skirtingų NT objektų vertės nustatymo metodų, kurie tradiciškai skirstomi į tris pagrindines kategorijas – generuojamų pajamų, kaštų ir rinkos palyginimo. Kaip teigia H. Usman ir kt. (2020) [11], kiekviena iš šių kategorijų praktikoje gali būti detalizuojama specifiniais metodais. Šie metodai remiasi skirtingomis prielaidomis apie vertinamą objektą bei jo naudą, tačiau kiekvienas iš jų turi tam tikrų problemų – pavyzdžiui, generuojamų pajamų metodai priklauso nuo

subjektyvių prognozių, o kaštų metodai dažnu atveju ignoruoja realias rinkos sąlygas. Rinkos palyginimo metodai šiuo metu sulaukia vis daugiau dėmesio dėl augančio prieinamų duomenų kiekio, šių metodų pagrindu vis dažniau vystomi duomenimis grįsti prognozavimo modeliai [11]. Viena iš tokių alternatyvų yra hedonistinės kainodaros metodas (*HPM*), kuris laikomas vienu iš labiausiai paplitusių būsto vertės nustatymo metodų tiek akademinėje, tiek praktinėje srityje. *HPM* grindžiamas prielaida, kad būsto vertę lemia ne vien bendra objekto būklė ar lokacija, bet ir įvairių požymių visuma – pradedant plotu, aukštu, statybos metais ir baigiant kaimynystės saugumu, atstumais iki įvairių įstaigų ir pan. [11]. Bendru atveju *HPM* leidžia išskaidyti objekto kainą į „implicitines kainas" (angl. *implicit prices*), kurios atskleidžia kiekvieno požymio įtaką. Tokiu būdu atsiranda galimybė vertinti, kurie požymiai geriausiai apibūdina objekto vertę. Vis dėlto, naudojant *HPM* susiduriama su metodologiniais iššūkiais. Pirmosios problemos atsiranda vertinant prielaidas apie normalumą, tiesinę priklausomybę ir homoskedastiškumą. Pasak H. Usman ir kt., klasikiniai regresijos modeliai remiasi prielaidomis, kad paklaidos yra suderintos su normaliuoju skirstiniu, kad tarp nepriklausomų kintamųjų ir priklausomojo kintamojo egzistuoja tiesinis ryšys, o paklaidų dispersija yra vienoda visose stebėjimų imtyse [11]. NT duomenyse šios prielaidos dažnai yra pažeidžiamos – pavyzdžiui, labai mažas ar labai didelis buto plotas gali turėti neproporcingą įtaką kainai. Kita dažnai pasitaikanti problema – multikolinearumas tarp nepriklausomų kintamųjų. Kaip teigia H. Usman ir kt., daugelis būsto charakteristikų yra glaudžiai susijusios tarpusavyje – pavyzdžiui, didesnio ploto butai paprastai turi daugiau kambarių, o naujos statybos butai dažnai pasižymi didesniu energiniu efektyvumu. Toks kintamųjų persidengimas gali apsunkinti kiekvieno požymio įtakos vertinimo procesą ir sumažinti modelių prognozių stabilumą [11]. Siekiant spręsti multikolinearumo problemą, tyrimų autoriai rekomenduoja taikyti pagrindinių komponenčių analizę (*PCA*) arba naudoti tik vieną iš koreliuojančių kintamųjų, tačiau net ir tokie veiksmai neišsprendžia visų problemų. Galiausiai, hedonistiniai metodai dažnai ignoruoja erdvinės priklausomybės poveikį. H. Usman ir kt. akcentuoja, kad būsto kainai labai daug įtakos turi aplinka – net ir labai panašūs objektai gali smarkiai skirtis kainomis vien dėl skirtingo mikrorajono [11]. Negana to, klasikinėje regresijoje neatsižvelgiama ir į tai, kad gretimi objektai daro įtaką vienas kito kainai (angl. *spatial autocorrelation*). Autoriai rekomenduoja taikyti rinkos segmentavimą arba naudoti erdvinius regresijos modelius, kurie geba įvertinti lokacinius požymius.

Tą pačią erdvinės priklausomybės problemą akcentuoja ir Y. Xiao ir kt. (2017) [12], analizuodami būsto kainas Pekine. Kaip pažymi autoriai, hedonistiniai metodai yra labai jautrūs erdvinei autokoreliacijai, tokiu būdu pažeidžiama esminė regresijos prielaida – stebinių nepriklausomumas. Šią problemą autoriai sprendė panaudodami patobulintą hedonistinės regresijos metodą su integruotu tikrinių vektorių filtravimu (angl. *Eigenvector Spatial Filtering*). *ESF* metodas papildo regresijos modelį erdviniais kintamaisiais – tikriniais vektoriais, kurie generuojami remiantis Morano *I* statistika, tokiu būdu įvertinamas erdvinės autokoreliacijos poveikis tarp stebimų reikšmių. Modelio konstravimas atliktas dviem etapais – pirmiausia iš visų galimų tikrinių vektorių atrinkti tie, kurie tenkina iš anksto nustatytą Morano *I* santykinio reikšmingumo slenkstį. Antrajame etape atlikta žingsninė regresija (angl. *stepwise regression*), kurios metu iš turimų tikrinių vektorių kandidatų ir 20 nepriklausomų kintamųjų (būsto ir kaimynystės charakteristikų) sudarytas galutinis modelis [12]. Analizuojant tyrimo rezultatus Y. Xiao ir kt. palygino klasikinį *OLS*[2] (angl. *Ordinary Least Squares*) regresijos modelį ir *ESF* (žr. 2 lentelę). Sąlyginai mažos *VIF* indekso reikšmės parodė, kad tiek *OLS*, tiek *ESF* modeliuose nebuvo reikšmingo multikolinearumo, o *Breusch-Pagan* testo rezultatai

[2] https://en.wikipedia.org/wiki/Ordinary_least_squares

neparodė reikšmingo heteroskedatiškumo [12]. Nors *Jarque-Bera* testas atskleidė, kad liekanos nėra suderintos su normaliuoju skirstiniu, autoriai pažymi, kad šios prielaidos netenkinimas neturi daug įtakos *OLS* modelio koeficientams, kai stebėjimų imtis yra didelė [12]. Didžiausias skirtumas tarp modelių atsiskleidė vertinant erdvinę autokoreliaciją. *OLS* modelio liekanose buvo nustatyta reikšminga erdvinė priklausomybė (Morano I = 0,2324, p < 0,001), tuo tarpu *ESF* modelio liekanose ši priklausomybė nebuvo aptikta (Morano I = -0,03, p > 0,1). Galima teigti, kad *ESF* modelis efektyviai pašalino erdvinį triukšmą iš hedonistinės regresijos modelio. Be to, *ESF* modelis paaiškino didesnę priklausomojo kintamojo dispersijos dalį – koreguotas $R^2$ koeficientas buvo didesnis, lyginant su *OLS* modeliu [12].

**2 lentelė.** Y. Xiao ir kt. tyrimo rezultatai [12]

| | | OLS | | | ESF | | |
|---|---|---|---|---|---|---|---|
| | | β | VIF | p reikšmė | β | VIF | p reikšmė |
| Struktūriniai požymiai | Plotas | 0.1765 | 3.614 | *** | 0.1736 | 3.9289 | *** |
| | Miegamųjų sk | 0.0213 | 2.7127 | *** | 0.0221 | 2.8425 | *** |
| | Kambariu sk. | 0.0291 | 1.6303 | *** | 0.0287 | 1.7267 | *** |
| | Langų orientacija | 0.0114 | 1.2112 | *** | 0.0098 | 1.2598 | *** |
| | Statvbos metai | −0.0251 | 2.0304 | *** | −0.0318 | 2.4164 | *** |
| Lokaciniai požymiai | Regionas | −0.0162 | 1.4022 | *** | −0.006 | 2.9418 | ** |
| | *Xicheng* | 0.2258 | 2.3692 | *** | 0.2237 | 4.6954 | *** |
| | *Dongcheng* | 0.184 | 2.0835 | *** | 0.1676 | 3.8533 | *** |
| | *Chaoyang* | 0.0282 | 2.7626 | *** | 0.0198 | 5.9633 | *** |
| | *Haidian* | 0.1179 | 3.4012 | *** | 0.1261 | 5.8366 | *** |
| | Autobusų stotelės | NA | NA | NA | −0.0059 | 2.4996 | *** |
| | Parduotuvės | 0.0044 | 2.3815 | * | 0.0126 | 4.3491 | *** |
| | Metro stotelės | NA | NA | NA | 0.0045 | 2.1238 | ** |
| | Stovėjimo aikštelė | NA | NA | NA | 0.0119 | 3.282 | *** |
| | Pradinė mokykla | 0.0037 | 1.5073 | * | 0.0083 | 2.468 | *** |
| | Prekybos centras | −0.0065 | 1.8605 | *** | −0.0126 | 3.4714 | *** |
| | Pramogų centras | 0.0096 | 1.3314 | *** | 0.0058 | 2.7566 | ** |
| | Sporto salė | 0.0061 | 1.7607 | *** | 0.0117 | 3.6287 | *** |
| Aplinkos požymiai | Oro tarša | −0.066 | 2.3913 | *** | −0.0601 | 5.7194 | *** |
| | Laiko požymis | 0.0162 | 1.0043 | *** | 0.0182 | 1.0146 | *** |
| Koreguotas R2 | | 0.8384 | | *** | 0.8768 | | *** |
| Liekanos | Breusch-Pagon | 0.5071 | | 0.4764 | 2.2747 | | 0.1315 |
| | Jarque-Bera | 3351.4 | | *** | 17021 | | *** |
| | Moran's I | 0.2324 | | *** | −0.03 | | 1 |

* Reikšmingas su p-reikšme = 0.05; ** Reikšmingas su p-reikšme = 0.01; *** Reikšmingas su p-reikšme = 0.001.

Analizuojant požymių įtaką, abu modeliai parodė, kad struktūrinės buto savybės (pvz., plotas, kambarių skaičius, langų orientacija) turi statistiškai reikšmingą teigiamą poveikį kainai – tai patvirtina teigiamos *beta* koeficientų reikšmės. Be struktūrinių ir lokacinių požymių, galutinio modelio rezultatai taip pat atskleidė ir aplinkos kokybės svarbumą – oro kokybės indeksas turėjo reikšmingą neigiamą poveikį buto kainai (*ESF beta* = -0,0601) [12]. Pasak tyrimo autorių, šis rezultatas sutampa su ankstesniais hedonistinių modelių tyrimais Čikagoje [13] ir Seule [14], kuriuose taip pat nustatytas neigiamas oro teršalų poveikis būsto kainoms.

Kaip pažymi S. Choi ir M. Y. Yi (2021) [15], būsto kainą galima suvokti kaip fizinių, lokacinių savybių ir makroekonominių veiksnių kombinaciją. Nors tradiciniai hedonistiniai metodai dažnai parodo gerus rezultatus, kai naudojami trumpo laikotarpio (iki pusės metų) duomenų rinkiniai, ilgesnio laikotarpio duomenų analizėje neišvengiamai išryškėja makroekonominių veiksnių poveikis [15]. Tai reiškia, kad ignoruojant šiuos veiksnius, modelių tikslumas gali reikšmingai sumažėti. Atsižvelgdami į šias problemas, S. Choi ir M. Y. Yi pasiūlė unikalią *HPM* alternatyvą – *Pseudo Self-Comparison Method*. *PSCM* esmė – kiekvienam stebėjimui surasti „pseudo dvynį“. Šis terminas apibrėžiamas kaip anksčiau parduotas būstas, kurio savybės labiausiai atitinka vertinamo objekto charakteristikas. Tada koreguojama surasto „pseudo dvynio“ sandorio kainą, atsižvelgiant į makroekonominių rodiklių pokyčius, kurie įvyko nuo sandorio datos iki vertinimo momento [15]. *PSCM* buvo išbandytas pasitelkiant didelės apimties NT sandorių duomenų rinkinį, apimantį Seulo

miestą ir Gyeonggi regioną – tai yra dvi svarbiausios Pietų Korėjos teritorijos, kuriose būsto rinka pasižymi dideliu pirkėjų aktyvumu ir kainų kintamumu [15]. Tyrime autoriai sukonstravo du scenarijus – „stabilų“ ir „kylantį“. Pirmuoju atveju modeliai buvo apmokomi su 2010–2017 metų duomenimis, o testavimas atliktas naudojant 2018 metų pirmojo pusmečio duomenis. Antruoju atveju apmokymas atliktas su 2010–2017 ir 2018 metų pirmojo pusmečio duomenimis, o testavimas atliktas naudojant 2018 metų antrojo pusmečio duomenis – šiuo laikotarpiu NT kainos sparčiai augo, tokiu būdu „kylantis“ scenarijus leido įvertinti modelių tikslumą esant besikeičiančioms rinkos sąlygoms.

Tam, kad būtų galima įgyvendinti *PSCM*, pradinis duomenų rinkinys buvo papildytas dviejų tipų požymių grupėmis – ankstesnio būsto sandorio charakteristikomis (*PP*) ir makroekonominių rodiklių pokyčiais (*MC*). *PP* požymiai siejami su „artimiausiu pseudo dvynio“ sandoriu ir jo data (*T-1*), kuri yra palyginama su vertinamo objekto sandorio data (*T0*). Norėdami tiksliau įvertinti „atstumą“ tarp šių dviejų laiko momentų, autoriai panaudojo normalizavimo transformaciją, kuri yra paremta išgyvenimo funkcija (angl. *Survival Function*) [15]. Tuo tarpu *MC* požymių grupę sudarė trys savaitiniai rodikliai, kuriuos pateikė viena didžiausių Pietų Korėjos finansinių institucijų. *KB* indeksas atspindėjo konkretaus rajono rinkos kainų pokytį (lyginant su 2015 metų bazinėmis reikšmėmis), *BS* indeksas parodė pirkėjų ir pardavėjų santykį, o *SS* – pardavėjų aktyvumą, lyginant su ankstesniu laikotarpiu [15]. Visgi ne visiems objektams pavyko priskirti „pseudo dvynius“, todėl dalis įrašų buvo pašalinti. Atlikus šį žingsnį, galutinis duomenų rinkinys buvo dar kartą patikrintas dėl multikolinearumo tarp neprikausomų kintamųjų. Panašiai kaip ir anksčiau nagrinėtuose tyrimuose, autoriai įvertino *PSCM* efektyvumą palygindami jį su klasikine *OLS* regresija, tačiau šį kartą papildomai išbandytos *Ridge*, *Lasso* ir *ElasticNet* reguliarizacijos formos. Reguliarizuotų modelių *alpha* hiperparametro paieškai naudotas „Grid Search“ algoritmas su atsitiktine atranka ir 5 kartų kryžminiu patikrinimu, o modelių vertinimui naudotas *MAPE* rodiklis. Tyrimo rezultatai rodė, kad *PSCM* pranoko *HPM* tiek stabilioje, tiek kylančioje rinkos situacijoje [15]. Naudodami *PSCM* ir atitinkamus *PP* ir *MC* požymių rinkinius, S. Choi ir M. Y. Yi gavo beveik penkis kartus mažesnes *MAPE* rodiklio reikšmes. Tokie rezultatai rodo, kad *PSCM* metodas yra itin efektyvus modeliuojant Pietų Korėjos NT kainas, kur dominuoja daugiabučiai pastatai, o didelė dalis butų pasižymi panašiomis charakteristikomis. Vis dėlto, reikėtų atkreipti dėmesį, kad šis metodas būtų sunkiau pritaikomas Europos šalyse, ypač Lietuvoje, kur butų įvairovė yra žymiai didesnė.

### 1.4. Sprendimų medžių ansamblių metodų taikymas NT vertei prognozuoti

Laikui bėgant, greta klasikinių regresinių metodų, būsto vertės nustatymo srityje pradėti taikyti ir pažangesni, mašininiu mokymusi paremti metodai. Kaip pažymi J. Hong ir kt. (2020) [16], naujų metodų paiešką iš dalies paskatino tarptautiniai įstatymai, tokie kaip Bazelio II susitarimas, kuriame numatyta, kad bankai privalo reguliariai pervertinti užstatytą turtą. Šios prievolės neišvengiamai padidina vertės nustatymo išlaidas – tiek laiko, tiek finansinių išteklių prasme, todėl atsiranda poreikis metodams, kurie gebėtų tiksliai ir greitai vertinti didelius kiekius NT objektų. Puikus pavyzdys – atsitiktinis miškas (angl. *Random Forest*), kuris pastaraisiais metais sulaukė nemažo dėmesio mokslinėje literatūroje. Šis modelis priskiriamas sprendimų medžių ansamblių klasei, kur prognozės formuojamos arba balsavimo principu (klasifikavimo uždaviniuose), arba skaičiuojant visų medžių prognozių vidurkį (regresijos uždaviniuose). Pagal E. A. Antipov ir kt. (2010) [17], vienas iš pagrindinių *RF* modelio privalumų yra paprastas konfigūravimas – rezultatus iš esmės lemia tik du hiperparametrai (bendras medžių skaičius ir maksimalus medžio gylis). Dėl savo paprastos struktūros ir nedidelio parametrų skaičiaus, *RF* modeliai yra aktualūs praktinėje veikloje, nes leidžia pasiekti didelį prognozavimo tikslumą neišnaudojant daug resursų.

J.Hong ir kt. tyrime *RF* modelis buvo išbandytas masinio būsto vertinimo (angl. *mass appraisal*) kontekste. Tyrimo duomenis sudarė Seulo Gangam rajono butų pardavimų įrašai, apimantys laikotarpį nuo 2006 iki 2017 metų [16]. Modeliavimui naudoti 26 kintamieji – struktūriniai, kaimynystės, lokaciniai ir makroekonominiai požymiai. Atlikę hiperparametrų optimizavimą, autoriai galiausiai pasirinko *RF* modelį, sudarytą iš 50 medžių, kurių maksimalus gylis buvo 17 [16]. Autoriai pabrėžia, kad nedideli tikslumo skirtumai tarp skirtingų *RF* modelių neatsveria papildomo skaičiavimo laiko, todėl pasirinktas modelis laikomas kompromisu tarp tikslumo ir naudojamų resursų [16]. Tolimesniame tyrimo etape autoriai atliko reikšmingų kintamųjų atranką, siekdami sumažinti modelio sudėtingumą bei išvengti galimo persimokymo. Atranka vykdyta eliminuojant mažiausiai reikšmingus požymius po vieną, o tikslumui vertinti naudotas *MAPE* rodiklis. Didžiausias tikslumas buvo pasiektas naudojant 16 kintamųjų iš 26 – tuomet *MAPE* reikšmė buvo mažiausia [16]. Atlikę požymių atranką, J. Hong ir kt. perėjo prie lyginamosios analizės, kurios metu buvo palygintas *RF* ir anksčiau minėtas klasikinis *OLS* regresijos modelis. Rezultatai atskleidė akivaizdų *RF* modelio pranašumą – visais vertintais laikotarpiais *RF* modelis generavo gerokai tikslesnes prognozes nei *OLS*. Autoriai taip pat pabrėžia, kad *RF* modelis pranoko *OLS* ne tik *MAPE* ir kitais rodikliais, bet ir rezultatų stabilumu – prognozės buvo mažiau jautrios treniravimo ir testavimo duomenų pasiskirstymams [16].

K. Teang ir Y. Lu (2021) [18] atliko panašios tematikos, medžių ansambliais paremtą tyrimą, tačiau jų darbe galima pastebėti keletą esminių skirtumų. Vienas iš jų – priklausomojo kintamojo pasirinkimas. Priešingai nei J. Hong ir kt. tyrime, K. Teang ir Y. Lu analizavo kvadratinio metro kainą, o ne bendrą buto vertę. Be to, siekdami įvertinti modelių elgseną esant skirtingiems duomenų pasiskirstymams, autoriai anksčiau minėtą *OLS* modelį pritaikė naudojant tiek originalų, tiek logaritmuotą priklausomą kintamąjį. Toks sprendimas leido įvertinti, ar priklausomojo kintamojo transformavimas pagerina prognozavimo rezultatus. Tyrime naudotą duomenų rinkinį sudarė 114 tūkstančių Stokholmo NT sandorių įrašų, apimančių laikotarpį nuo 2005 iki 2014 metų. Toks didelis rinkinys leido tyrimo autoriams įvertinti skirtingas duomenų išskaidymo strategijas [18]. Palyginimui buvo taikoma tiek atsitiktinė atranka, tiek „slenkančio lango“ (angl. *rolling window*) principu paremta strategija, kurioje modeliai apmokomi naudojant ankstesnio laikotarpio duomenis, o tikslumo vertinimas atliekamas naudojant vėlesnio laikotarpio duomenis. Tokiu būdu buvo analizuotas ne tik bendras modelių tikslumas, bet ir modelių atsparumas makroekonominių sąlygų pokyčiams [18]. Dar vienas svarbus skirtumas, lyginant su J. Hong ir kt. tyrimu, yra modelio hiperparametrai – K. Teang ir Y. Lu naudojo 500 ir 1000 medžių modelius. Pasirinkimas naudoti tokį didelį medžių kiekį leido autoriams padidinti prognozavimo tikslumą ir sumažinti atsitiktinių svyravimų įtaką rezultatams [18]. Remiantis tyrimo rezultatais, dar kartą įrodyta, kad *RF* modelis pranoksta abi *OLS* versijas. Nors treniravimo imties geriausia *MAPE* rodiklio reikšmė gauta naudojant paprastąjį *OLS* modelį, testavimo etape atsitiktinio miško modelis aiškiai dominavo pagal visus kokybės rodiklius.

Q. Truong ir kt. (2020) [19] atliko platesnę medžių ansambliais grįstų modelių analizę panaudojant ne tik *RF*, bet ir tokius algoritmus kaip *XGBoost*, *LightGBM* bei kelių modelių derinius, paremtus *Stacked Generalization* metodika. Kaip pažymi autoriai, tokia metodų įvairovė grindžiama tuo, kad būsto kainų indeksas tėra apytikslis rodiklis, apskaičiuojamas iš visų sandorių duomenų, todėl jis nėra tinkamas prognozuoti konkretaus NT objekto vertę [19]. Autoriai panaudojo „Housing Price in Beijing“ duomenų rinkinį[3], apimantį daugiau nei 300 tūkstančių NT sandorių, įvykusių 2009–2018

[3] https://www.kaggle.com/datasets/ruiqurm/lianjia

metais. Prieš pradedant modeliavimą, autoriai atliko duomenų valymą ir žvalgomąją analizę, sutvarkytą duomenų rinkinį sudarė 19 kintamųjų – 9 kiekybiniai ir 10 kategorinių. Atlikus žvalgomąją analizę, pastebėta, jog egzistuoja stiprios koreliacijos tarp buto kv. metro pardavimo kainos ir atstumo iki miesto centro, pastato amžiaus požymių [19].

Modeliavimo etape Q. Truong ir kt. išbandė penkis skirtingus mašininio mokymosi algoritmus – atsitiktinį mišką, ekstremalų gradientinį pastiprinimą (*XGBoost*), švelnią gradientinio didinimo mašiną (*LightGBM*), hibridinę regresiją (*Hybrid regression*) ir sudėtinį *Stacked Generalisation* modelį. *RF*, *XGBoost* ir *LightGBM* algoritmams atliktas hiperparametrų optimizavimas, naudojant *GridSearchCV*[4] funkciją iš *scikit-learn* bibliotekos. Tuo tarpu hibridinei regresijai ir sudėtiniam modeliui atskiras derinimas nebuvo atliktas, kadangi šie metodai remiasi ankstesniame etape optimizuotais modeliais. Nors tiek hibridinė regresija, tiek sudėtinis modelis išnaudoja kelių regresinių modelių sąveiką, jų veikimo principai, pasak autorių, skiriasi. Hibridinė regresija paremta paprastu prognozių vidurkinimu, o *Stacked Generalisation* pasižymi papildomu mokymosi etapu, kai iš pirminio modelio prognozių liekanų sukuriamas dar vienas modelis. Taip pat reikėtų atkreipti dėmesį, kad tyrimo rezultatai vertinti naudojant *RMSLE* (angl. *Root Mean Squared Logarithmic Error*) rodiklį. Šis rodiklis leidžia įvertinti santykines modelių prognozių paklaidas, todėl jis yra tinkamas taikyti NT rinkos kontekste, kur duomenų imtys dažnai pasižymi didele priklausomojo kintamojo dispersija [19]. Q. Truong ir kt. tyrimo rezultatai atskleidė, kad kiekvienas iš penkių algoritmų pasižymėjo skirtingais privalumais ir trūkumais. Geriausi rezultatai treniravimo imtyje pasiekti naudojant *RF* modelį, tačiau testavimo imtyje, kur svarbiausias yra bendras modelio pajėgumas, išsiskyrė *Stacked Generalisation* modelis, kurio *RMSLE* reikšmė buvo mažiausia. Autoriai taip pat vizualiai palygino hibridinę regresiją ir sudėtinį modelį. Pastebėta, jog hibridinė regresija tiksliau prognozavo kraštutines (labai dideles arba labai mažas) butų kainas, o *Stacked Generalisation* pasižymėjo didesniu tikslumu prognozuojant vidutinės vertės butus. Galima susidaryti nuomonę, kad sudėtinis modelis yra šiek tiek pranašesnis, kai tikslas – gauti stabilesnes prognozes [19]. Autoriai taip pat atkreipia dėmesį į praktines problemas – nors hibridinė regresija ir sudėtinis modelis demonstravo didelį prognozių tikslumą, šie metodai pasižymi ilgai trunkančiais skaičiavimo procesais, daugiausia dėl *RF* komponento ir kryžminio patikrinimo procedūrų. Dėl šių priežasčių, tolimesniuose tyrimuose autoriai rekomenduoja ieškoti efektyvesnių modelių derinių, gilintis į veiksnius, lemiančius medžių ansamblių modelių pajėgumą bei tirti mašininio ir giluminio mokymosi (angl. *deep learning*) modelių derinius [19].

C. Kmen ir kt. (2024) [20] akcentavo *XGBoost* algoritmo efektyvumą prognozuojant NT kainas. Šiame tyrime naudotas dešimties metų laikotarpio Vienos miesto NT duomenų rinkinys, kuriame buvo 200 skirtingų požymių. Panašiai kaip ir [18] tyrime, tokios didelės apimties duomenų rinkinys leido autoriams išbandyti skirtingas modelių apmokymo strategijas. Priklausomas kintamasis šiame tyrime buvo buto kvadratinio metro kaina, o vertindami skirtingų specifikacijų *XGBoost* modelius autoriai panaudojo *MAPE* rodiklį. Autoriai taip pat priėme sprendimą suskirstyti duomenis pagal buto statybos metus. Rezultatai atskleidė, kad modeliai, treniruoti su naujos statybos butų duomenų rinkiniais, pasiekė reikšmingai geresnius rezultatus – *MAPE* rodiklio reikšmė buvo vidutiniškai 6% mažesnė, lyginant su modeliais, kuriuose naudojamas bazinis duomenų rinkinys. Be to, tyrimas parodė, kad atliekant prognozavimą su ilgesnio laikotarpio (iki 3 metų) duomenų rinkiniais, modelių tikslumas išliko gana didelis – *MAPE* rodiklio reikšmė neviršijo 20% [20]. Taip pat svarbu paminėti,

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

kad daug įtakos *XGBoost* modelio rezultatams turėjo papildomai įtraukti erdviniai bei demografiniai požymiai, kurie leido tiksliau įvertinti aplinkos įtaką buto kvadratinio metro kainai.

### 1.5. Kiti tyrimai

J. Kalliola, J. Kapočiūtė-Dzikienė ir R. Damaševičius (2021) [21] taip pat nagrinėjo būsto vertės nustatymo problemą, šiame tyrime pagrindinis dėmesys skirtas dirbtiniems neuroniniams tinklams (angl. *Artificial Neural Networks*). Duomenys surinkti iš skirtingų Suomijos nekilnojamojo turto portalų, kuriuose pateikiama informacija apie būsto sandorius Helsinkyje, galutinį duomenų rinkinį sudarė 43 požymiai. Modeliavimui autoriai pasirinko daugiasluoksnį perceptroną (angl. *Multilayer Perceptron*) – vieną iš populiariausių dirbtinių neuroninių tinklų architektūrų, kuri yra plačiai taikoma regresijos uždaviniuose. Siekiant pagerinti modelių prognozavimo rezultatus, tyrimo autoriai atliko išsamų hiperparametrų optimizavimą, naudojant „Weights & Biases" įrankį su Bajeso paieškos algoritmu [21]. Optimizavimo metu automatiškai koreguotas partijos dydis, sluoksnių skaičius, mokymosi greitis ir *dropout* parametro reikšmė. Visi modeliai buvo vertinami pagal vidutinės kvadratinės paklaidos (*MSE*) rodiklį, tai leido identifikuoti optimalų hiperparametrų derinį. Geriausią rezultatą pasiekęs modelis sukurtas naudojant šiuos hiperparametrus: aktyvacijos funkcija – *ReLU*, optimizavimo algortimas – *Adam*, partijos dydis – 550, *dropout* reikšmė – 0,005, mokymosi greitis – 0,0012 bei validacijos imties dydis – 8% [21]. Optimizavus modelio hiperparametrus buvo dar kartą pagerintas bendras prognozavimo tikslumas – determinacijos koeficiento reikšmė padidėjo per 0,05, o *MSE* reikšmė sumažėjo net 24,7% [21]. Galima teigti, kad modelių hiperparametrų optimizavimas yra itin svarbi procedūra NT vertinimo uždaviniuose. Panašią išvadą pateikia ir A. Deaconu ir kt. (2022) [22], kurie savo tyrime palygino dirbtinius neuroninius tinklus ir apibendrintą tiesinį modelį (angl. *Generalized linear model*). Tyrimui naudoti Klužo-Napokos miesto butų sandorių duomenis, o modelių vertinimui naudoti 5 skirtingi kokybės kriterijai. Visi jie parodė aiškų *ANN* modelio pranašumą. Atlikus hiperparametrų optimizavimą, *RMSE* rodiklio reikšmė sumažėjo apie 12%, o determinacijos koeficiento reikšmė padidėjo apie 7,5% [22]. Tokie rezultatai leidžia susidaryti nuomonę, kad neuroninių tinklų modeliai pakankamai gerai prisitaiko prie sudėtingų duomenų struktūrų ir užtikrina tikslias prognozes, nepaisant to, kad apmokymui naudojami sąlyginai maži duomenų rinkiniai. A. Deaconu ir kt. taip pat akcentuoja, kad praktikoje šie modeliai gali būti taikomi ne tik prognozių generavimui, bet ir papildomai NT vertinimo kokybei užtikrinti – pavyzdžiui, identifikuoti įtartinus, reikšmingai nuo rinkos vidurkio nukrypstančius objektus. Vis dėlto, J. Kalliola ir kt. pabrėžia esminę *ANN* problemą – neuroniniai tinklai pasižymi menku paaiškinamumu, jų veikimo principai yra sunkiai interpretuojami, todėl šie modeliai yra dažnai kritikuojami dėl „juodosios dėžės"[5] efekto [21].

Taip pat verta paminėti T. Potrawa ir A. Tetereva (2022) [23] atliktą tyrimą, kuriame panaudoti ne tik pažangūs mašininio mokymosi metodai, bet ir unikalūs duomenų šaltiniai. Skirtingai nei daugelyje ankstesnių tyrimų, kuriuose daugiausia naudoti struktūriniai, lokaciniai ar aplinkos požymiai, šiame darbe stengtasi įtraukti ir nestruktūrizuotus duomenis – būsto nuotraukas bei aprašymus. Įdomu ir tai, kad tyrimas atliktas naudojant nuomos rinkos duomenis. Pasak tyrimo autorių, toks sprendimas leido išvengti derybų įtakos kainai, tokiu būdu tyrimo rezultatai buvo patikimesni [23]. Atliekant tyrimą buvo automatizuotas duomenų rinkimo procesas, kurio metu iš įvairių Nyderlandų skelbimų portalų buvo surinkti Roterdamo mieste nuomojamų butų duomenys. Duomenų rinkinį, kuriam surinkti prireikė 2 mėnesių, sudarė 1844 unikalūs įrašai, kartu su pagrindiniais požymiais buvo sukaupta apie

[5] https://en.wikipedia.org/wiki/Black_box

40 tūkstančių nuotraukų [23]. Tekstinių duomenų analizė autoriams leido išskirti 3 papildomus kategorinius kintamuosius, susijusius su leidimu laikyti gyvūnus, būsto insoliacija bei privaloma pajamų deklaracija. Tuo tarpu vaizdinė medžiaga buvo analizuojama pasitelkiant konvoliucinius neuroninius tinklus, kurių pagalba iš analizuotų nuotraukų turinio išgauti dar 2 požymiai – kategoriniai kintamieji, nurodantys, ar butas turi vaizdą į žalumą arba į vandens telkinius [23]. Panašiai kaip ir [16], [18] tyrimuose, šiame darbe autoriai palygino *OLS* ir *RF* modelius, tačiau šį kartą, vertinant modelių tikslumą, autoriai naudojo skirtingus duomenų rinkinius. Iš pradžių modeliai buvo testuojami naudojant tik pagrindinius struktūrinius požymius ir objekto koordinates, vėliau įtrauktas požymis, nurodantis kelionės iki miesto centro laiką, o į paskutinį duomenų rinkinį įtraukti tekstinių ir vaizdinių duomenų pagalba sukurti požymiai. Rezultatai rodė, kad *RF* modelis visuose duomenų rinkiniuose pranoko *OLS* tiek pagal determinacijos koeficiento reikšmę, tiek pagal *RMSE* reikšmę. Tai dar kartą patvirtina sprendimų medžių ansambliais paremtų modelių pranašumą prognozuojant kintamuosius, kurie turi sudėtingas, netiesines priklausomybes. Taip pat išsiaiškinta, kad įtraukus tekstinių ir vaizdinių duomenų požymius, galima padidinti sprendimų medžių ansamblių modelių tikslumą – naudojant papildytą duomenų rinkinį *RF* modelio *RMSE* rodiklio reikšmė apie 2,4% [23].

C. M. Caroni (2022) [24] taip pat bandė prognozuoti nuomos kainas, tačiau šį kartą vertinta, kaip globalūs modeliai, apmokyti visos šalies mastu, skiriasi nuo lokalių modelių. Tyrime naudota keletas anksčiau neminėtų modelių – *ExtraTree* bei apibendrintas adityvinis modelis (angl. *Generalized additive model*) [24]. Pastebėta, kad lokalūs modeliai, apmokyti naudojant konkretaus miesto duomenis, pranoko globalius modelius pagal visus vertinimo kriterijus. Kaip nurodo autorė, Paryžiaus duomenų atveju, naudojant lokalų *XGBoost* modelį, *MAPE* ir *MDAPE* reiškmės sumažėjo atitinkamai 0,43% ir 0,33% [24]. Nors tai nėra didelis tikslumo pokytis, galima susidaryti nuomonę, kad rinkos segmentavimas yra svarbus prognozuojant tiek buto pardavimo, tiek nuomos kainas. C. M. Caroni taip pat pažymi, kad globalūs modeliai gali būti naudingi, kai trūksta duomenų apie tam tikras vietoves [24].

Iš naujesnių tyrimų verta išskirti H. Sharma ir kt. (2024) [25] darbą, kuriame būsto vertės nustatymo problema spręsta taikant visus anksčiau paminėtus metodus – klasikinius tiesinės regresijos, sprendimų medžių ansamblių ir neuroninių tinklų modelius. Tyrime naudotas viešai prieinamas Ames miesto NT duomenų rinkinys[6], kurį sudarė 2900 įrašų ir 82 paaiškinamieji požymiai, priklausomas kintamasis šiuo atveju buvo namo pardavimo kaina (*SalePrice*). Autoriai atliko penkių algoritmų palyginimą, modeliai buvo treniruoti taikant tą pačia duomenų išskaidymo strategiją ir kryžminį patikrinimą. Nemažai dėmesio skirta hiperparametrų optimizacijai, panašiai kaip ir [19] tyrime, naudota *GridSearchCV* funkcija. Ypatingas tikslumo padidėjimas užfiksuotas *MLP* modelyje – atlikus optimizavimo procedūrą koreguoto determinacijos koeficiento reikšmė padidėjo beveik 27%, kas akivaizdžiai rodo *MLP* jautrumą parinktiems hiperparametrams [25]. Geriausius rezultatus eilinį kartą demonstravo *XGBoost*, pranokęs kitus modelius pagal visus naudotus vertinimo kriterijus. Atsižvelgdami į rezultatus, autoriai būtent šį modelį pasirinko atliekant tolimesnę paaiškinamųjų požymių svarbumo analizę. Nustatyta, kad daugiausia įtakos namo kainai turi bendras kokybės įvertinimas, pirmojo aukšto gyvenamasis plotas, garažų vietų skaičius bei rūsio plotas [25]. Šios įžvalgos gali būti naudingos praktikoje – tiek NT vystytojams, tiek būsimiems pirkėjams, siekiantiems pasirinkti optimalią alternatyvą. Autoriai taip pat akcentuoja, kad didelio duomenų

[6] https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset

kiekio apdorojimas reikalavo nemažai resursų – net kai buvo naudota mažesnė nei trijų tūkstančių stebėjimų imtis, hiperparametrų optimizavimo procedūra neuroninių tinklų modeliui užtruko apie 8 valandas, todėl, dirbant su didesniais duomenų rinkiniais, autoriai rekomenduoja ieškoti efektyvesnių modeliavimo sprendimų [25].

## 1.6. Literatūros analizės išvados

Literatūros apžvalgoje stengtasi nagrinėti būsto vertės nustatymo problemą pagal skirtingus aspektus. Vienas iš jų – būsto kainų prognozavimas kaip laiko eilučių uždavinys, daugiausia dėmesio skiriant makroekonominių veiksnių įtakai būsto kainų indeksams. Šių indeksų analizė yra naudinga, kai siekiama įvertinti bendras nekilnojamojo turto rinkos tendencijas ar identifikuoti galimo „burbulo“ susiformavimo riziką. Vis dėlto, galima teigti, kad norint atsakyti į visus investuotojui kylančius klausimus, neužtenka vien indeksais ir laiko eilutėmis paremtos analizės. Negana to, net jeigu turėtumėme pakankamai duomenų apie būsto kainų indeksus pagal rajonus ar pašto kodus, išlieka problema, jog toje pačioje aplinkoje egzistuoja objektų įvairovė – nuo senų daugiabučių iki modernių apartamentų, nuo mažų studijų iki prabangių šeimyninių kotedžų. Visa tai leidžia susidaryti nuomonę, kad greta bendrų rinkos tendencijų analizės būtina nagrinėti ir kitą problemos suvokimą – būsto vertę pagal regresijos principus.

Atliekant literatūros apžvalgą pastebėta, jog daug dėmesio sulaukia hedonistiniai kainodaros metodai. Vis dėlto, taikant klasikinius daugialypės regresijos modelius, susiduriama su tam tikrais iššūkiais – dažnai pažeidžiamos bazinės prielaidos, o tarp paaiškinamųjų požymių pasitaiko multikolinearumas. Siekiant spręsti šias problemas, tyrimų autoriai siūlo taikyti reikšmingų požymių atranką arba atlikti rinkos segmentavimą ir tada pereiti prie modeliavimo proceso. Taip pat pastebėta, kad naudojant tokius požymius kaip rajoną, koordinates, atstumus iki miesto centro bei kitų svarbių lokacijų, galima dar labiau pagerinti modelių rezultatus. Be to, analizuojant tyrimus, kuriuose naudoti ilgesnio laikotarpio duomenų rinkiniai, pastebėta, kad tik struktūrinių ir lokacinių požymių nepakanka – reikėtų į modeliavimo procedūrą įtraukti papildomus makroekonominius kintamuosius.

Išnagrinėjus naujesnius tyrimus, pastebėta aiški tendencija, kai pereinama prie pažangesnių mašininio mokymosi modelių. Šie modeliai laikomi alternatyva klasikinės regresijos modeliams ir leidžia įvertinti sudėtingesnes, netiesines priklausomybes tarp būsto savybių ir jo kainos. 3 lentelėje pateikiami tyrimų rezultatai, kuriuose būsto kaina nagrinėjama kaip regresijos uždavinys. Perėjus prie mašininio mokymosi metodų analizės, pastebėta, kad *OLS* modelis vis tiek išlieka svarbus – daugelyje tyrimų jis panaudotas kaip „benchmark“ modelis, leidęs objektyviai įvertinti pažangesnių algoritmų pranašumus. Analizuotų tyrimų rezultatai parodė, kad didžiausiu prognozavimo tikslumu dažniausiai pasižymėjo atsitiktinio miško algoritmas, tačiau negalima pamiršti ir kitų sprendimų medžių ansambliais grįstų metodų, kurie taip pat demonstravo gerus rezultatus. Šių metodų populiarumui daug įtakos turi ir jų prieinamumas – daugelis algoritmų yra nesunkiai įgyvendinami naudojantis atviro kodo Python programavimo kalbos biblioteka *scikit-learn*[7].

Vertinant visus nagrinėtus tyrimus ir jų rezultatus, galima teigti, kad pasirinkimas naudoti vieną modelį su fiksuotais parametrais retai garantuoja geriausius rezultatus. Norint pasiekti maksimalų tikslumą, reikėtų ne tik išbandyti kelių modelių derinius bei jų prognozių vidurkį, bet ir skirti dėmesio kiekvieno modelio hiperparametrų optimizavimui. Tyrimų rezultatai parodė, kad tie patys algoritmai, priklausomai nuo parinktų hiperparametrų, gali pasižymėti skirtingu prognozių tikslumu. Taip pat

[7] https://scikit-learn.org/stable/

svarbu tinkamai pasirinkti priklausomąjį kintamąjį – kai kuriais atvejais prognozuojant kvadratinio metro kainą galima gauti tikslesnes prognozes, ypač kai duomenų rinkiniai pasižymi didele objektų įvairove.

Svarbu ir tai, kad optimalaus NT objekto pasirinkimas paremtas ne tik statistiškai pagrįsta kaina, bet ir investiciniu atsiperkamumu. Dėl šios priežasties šiame darbe pasirinkta nagrinėti ne tik pardavimo, bet ir nuomos rinkos duomenis, kurie sudarytų galimybę įvertinti būsto pardavimo–nuomos kainų santykį (angl. *sell-rent ratio*).

**3 lentelė.** Literatūros apžvalgos rezultatų suvestinė

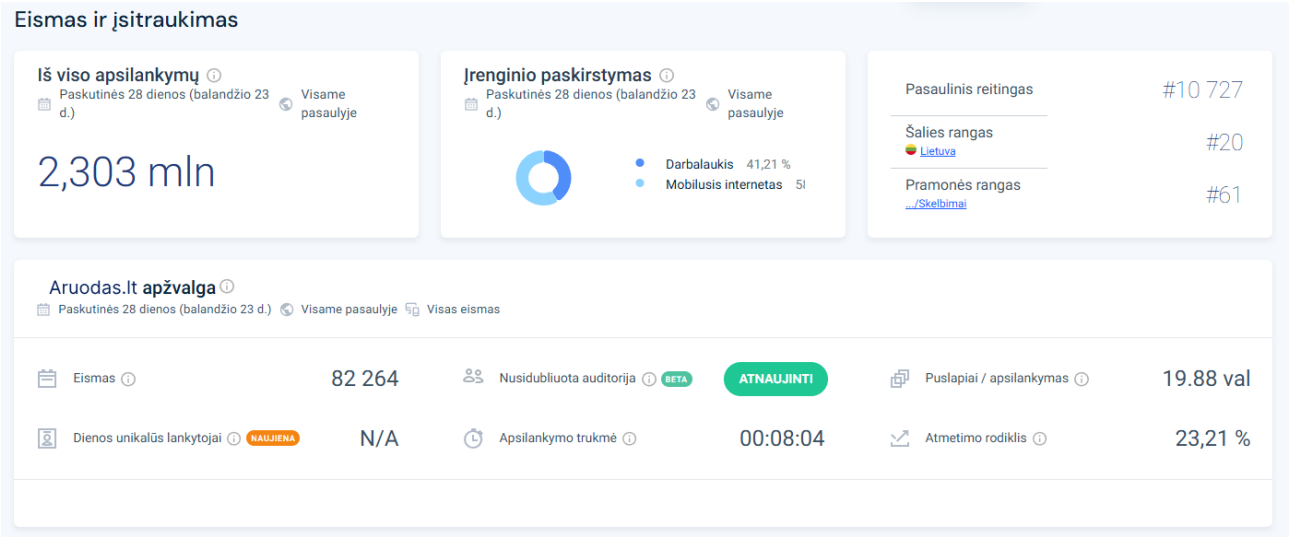
| **Straipnis** | **Duomenų rinkinio stebėjimų skaičius/paaiškinamųjų požymių skaičius** | **Naudoti algoritmai** | **Geriausias rezultatas** |
|---|---|---|---|
| [12] | 6959/21 | ESF | R2=0,8768 |
| [15] | 1,66 mln./27 | Lasso, Ridge, ElasticNet, PSCM | R2=0,965;MAPE=0,0587 |
| [16] | 16601/26 | RF | R2=0,976;MAPE=0,0542 |
| [18] | 114293/10 | RF | RMSE=7096,53;MAPE=0,1771;R2=0,8316 |
| [19] | 231962/19 | RF, XGBoost, LightGBM, Hybrid regression, Stacked Generalisation | RMSLE=0,16350 |
| [20] | 83527/200 | XGBoost | MAPE=0,1408 |
| [21] | 4041/43 | MLP | R2=0,95;RMSE=46815,7;RME=8,3%;MAE=23320,9 |
| [22] | 900/33 | GLM, ANN | MSE=572,28;RMSE=17162,74;RMSPE=1,88%;R2=0,8051 |
| [23] | 1844/21 | RF, CNN | R2=0,74;RMSE=240 |
| [24] | 314345/30 | GAM, ExtraTree, RF, XGBoost, ElasticNet | MAPE=0,0905;MDAPE=0,0655 |
| [25] | 2930/82 | RF, MLP, SVR, XGBoost | R2=0,92;MSE=0.015;RMSE=0,112;MAE=0,084 |

## 2. Tyrimo metodai

Atliekant šį tyrimą ir identifikuojant geriausius problemos sprendimo metodus, nemažai dėmesio skirta ir S. Adomavičiaus (2022) [26] darbui, kuriame modeliuojant butų kainas buvo taikytos įvairios duomenų analizės ir mašininio mokymosi technikos. Šiame tyrime buvo remiamasi kai kuriais S. Adomavičiaus metodologiniais sprendimais, kurie yra susiję su duomenų paruošimu, požymių inžinerija bei modelių optimizavimu, keletas pavyzdžių būtų tokios Python bibliotekos kaip *Boruta*, *Optuna* ir *Yellowbrick*. Nors tyrimo metu buvo ieškoma alternatyvių metodų, tačiau jų pritaikymas dažnu atveju nesuteikė geresnių tikslumo rezultatų, lyginant su tais, kurie buvo gauti minėtame darbe. Dėl šios priežasties buvo pasirinkta grįžti prie kai kurių S. Adomavičiaus metodologinių sprendimų ir adaptuoti juos prie šio darbo. Svarbu pažymėti, kad naudoti metodai nebuvo aklai perimti – jie buvo pagrįsti perbandymais, koreguoti ir taikyti savarankiškai atsižvelgiant į skirtingas duomenų specifikacijas bei išsikeltus uždavinius.

### 2.1. Viešųjų duomenų rinkimas

Atsižvelgiant į didelį NT objektų kiekį ir nuolatinį skelbimų atsinaujinimą, galima teigti, kad duomenų rinkimas rankiniu būdu yra ypatingai neefektyvus. Siekiant padidinti efektyvumą laiko atžvilgiu, duomenų rinkimo procesas buvo automatizuotas, panaudojant programinę įrangą. Duomenys buvo renkami iš internetinės svetainės „aruodas.lt“, kurioje pateikta Lietuvos būsto skelbimų informacija. Svarbu pabrėžti, kad bet kokį automatizuotą duomenų rinkimo procesą reikėtų vykdyti atsakingai – vadovaujantis galiojančiais teisiniais reglamentais, etikos principais bei techninėmis rekomendacijomis. Svarbiausias aspektas – užtikrinti, kad duomenų rinkimas nesukeltų perteklinio krūvio svetainės serveriams. Detaliau panagrinėjus šį aspektą, išsiaiškinta, kad „aruodas.lt“ svetainėje nėra aiškių duomenų rinkimo gairių ar apribojimų, tačiau tai nereiškia, kad automatizuotas procesas gali būti vykdomas neapgalvotai – prieš tai būtina įvertinti svetainės statistinę informaciją ir nustatyti, ar siunčiamų užklausų kiekis nesukels problemų svetainės funkcionalumui. Remiantis analitinės platformos *SimilarWeb* duomenimis (žr. 4 pav.) [27], galima pastebėti, kad svetainėje kiekvieną dieną užfiksuojama apie 70-80 tūkstančių apsilankymų, o vienas apsilankymas vidutiniškai užtrunka 8 minutes, taip pat galima matyti, kad vieno apsilankymo metu yra atidaroma beveik 20 skirtingų puslapių.




**4 pav.** Svetainės „aruodas.lt“ statistinė informacija (žiūrėta 2025-04-23) [27]

Atsižvelgiant į šią statistinę informaciją, galima teigti, kad ribotos apimties duomenų rinkimas, orientuotas į konkretų svetainės segmentą, pavyzdžiui, butų pardavimo skelbimus Vilniuje arba nuomos skelbimus Kaune, neturi reikšmingo neigiamo poveikio svetainės funkcionalumui.

Automatizuotam duomenų rinkimui panaudotos dvi Python programavimo kalbos bibliotekos – *Selenium* ir *BeautifulSoup*. Naudojamos kartu, šios priemonės sudarė galimybes pasiekti ir apdoroti tiek *HTML* bei *XML* turinį, tiek dinaminiu būdu (per *JavaScript*) generuojamą informaciją, kuri, naudojant įprastas priemones, būtų nepasiekiama. *Selenium* biblioteka, kartu su *WebDriver* įrankiu, padėjo imituoti naudotojo naršymą ir leido išgauti pilnai sugeneruotą svetainės turinį, o *BeautifulSoup* biblioteka užtikrino duomenų struktūrizavimą ir reikalingų elementų išskyrimą iš *HTML* failo. Pirmasis duomenų rinkimo etapas buvo vykdomas naudojant atskirą Python kalbos kodą. Ši sukonstruota programa, naudodama *Selenium* biblioteką ir „headless“ režimu veikiančią naršyklę, automatiškai naršė po svetainės puslapius ir rinko puslapiuose esančias *URL* formato nuorodas. Siekiant išvengti pasikartojimų, surinktos nuorodos buvo talpinamos *set* tipo struktūroje, o į galutinį *CSV* formato failą įrašomos tik unikalios reikšmės. Duomenų rinkimas atliktas 2025 metų kovo 14 – balandžio 14 dienomis. Antrajame etape buvo naudojamas atskiras kodas, kurio pagalba iš rinktų *URL* nuorodų išgauta struktūrizuota informacija apie kiekvieną skelbimą. Kodo veikimo logika gali būti suskirstyta į kelis pagrindinius etapus:

1. Duomenų paruošimas ir puslapio apdorojimas:

   – nuskaitomas failas, kuriame saugomos skelbimų nuorodos pagal rajonus, tikrinama, kurių rajonų duomenys jau buvo apdoroti;
   – sukuriamas failas, į kurį bus įrašomi visi nuskaityti duomenys;
   – naudojant *Selenium* ir *WebDriver*, po vieną atidaromos skelbimų nuorodos, automatiškai priimami slapukai;
   – naudojant *BeautifulSoup*, nuskaitomas visas *HTML* turinys;

2. Pagrindinės objekto charakteristikos:

   – iš *HTML* turinio surenkama pagrindinė informacija apie objektą – kaina, plotas, kambarių skaičius, aukštas, aukštų skaičius, statybos metai ir pan.;
   – naudojant skelbimo apačioje esantį aprašymą, papildomai patikrinama, ar yra tokios ypatybės kaip balkonas, rūsys bei vieta automobiliui;
   – visi duomenys grupuojami pagal *HTML* struktūroje esančias reikšmes ir priskiriami atitinkamiems kintamiesiems.

3. Geografinė informacija ir aplinkos duomenys:

   – automatiškai paspaudžiamas puslapyje esantis interaktyvus žemėlapis, iš naujai atsidariusio lango nuskaitomos tikslios objekto koordinatės;
   – surenkami duomenys apie atstumus iki artimiausių darželių, mokyklų, parduotuvių ir viešojo transporto stotelių;
   – papildomai surenkami duomenys apie atstumus (važiuojant automobiliu) iki kitų reikšmingų vietų, pavyzdžiui, iki geležinkelio stoties;
   – tikrinama, ar skelbime yra pateikiami oro taršos duomenys – jei šie duomenys egzistuoja, tada nuskaitomos azoto dioksido ir kietųjų dalelių koncentracijų reikšmės.

Siekiant riboti siunčiamų užklausų kiekį ir neviršyti priimtino serverių apkrovos lygio, kiekvienos nuorodos apdorojimui skiriama 10 sekundžių. Jei duomenų nuskaitymas įvyksta greičiau, programa priverstinai sustabdoma, naudojant „time.sleep“ funkciją. Tokiu būdu siekiama išlaikyti vienodą laiko tarpą tarp užklausų ir imituoti natūralią vartotojo elgseną. Visi surinkti požymiai jungiami į vieną struktūrizuotą žodyną, duomenys įrašomi į atitinkamus *CSV* formato failus, kurie vėliau bus naudojami modeliavimo etape. Duomenų rinkimo etape atsisakyta rinkti jautrią informaciją, tokią kaip skelbimų autorių ar brokerių vardai, kontaktiniai duomenys ir pan. Šis sprendimas leido užtikrinti asmens duomenų apsaugą bei išlaikyti konfidencialumą.

## 2.2. Žvalgomosios duomenų analizės įrankiai

Šioje tyrimo dalyje panaudoti keli pagrindiniai Python programavimo kalbos įrankiai, kurie leido efektyviai apdoroti ir analizuoti turimus duomenis. *Pandas* biblioteka leido nuskaityti turimus *CSV* formato failus bei atlikti pradines požymių transformacijas, o duomenų vizualizacijoms panaudota *matplotlib* biblioteka. Interaktyvių žemėlapių kūrimui panaudota *folium* biblioteka, šis įrankis leido geografiškai vizualizuoti NT objektų pasiskirstymus pagal jų koordinates, tokiu būdu galima lengviau atpažinti erdvinius dėsningumus ir įvertinti regioninius kainų skirtumus. Duomenų valymui panaudotos *re* modulio funkcijos, kurios leido suvienodinti atstumų požymių skales bei sutvarkyti kategorinių požymių pavadinimus. Kategorinių požymių kodavimui (pvz., šildymo tipui, pastato tipui ir pan.) taikyti *MultiLabelBinarizer*[8] ir *OneHotEncoder*[9] iš *sklearn.preprocessing* modulio – šie įrankiai leido išskleisti kategorinius požymius, paverčiant juos į binarinius „one-hot“ formato stulpelius.

## 2.3. Tiesinės regresijos modeliai

Nepaisant sparčiai augančio sudėtingų mašininio mokymosi algoritmų populiarumo, klasikiniai tiesinės regresijos modeliai vis dar išlieka vieni plačiausiai taikomų modeliavimo įrankių. Literatūros apžvalgoje minėta, jog paprasta daugialypė tiesinė regresija grindžiama prielaida, kad tarp priklausomojo kintamojo ir nepriklausomų kintamųjų egzistuoja tiesinis ryšys, kuris matematiškai gali būti išreikštas tokiu būdu:

$$y_i = \beta_0 + B'x_i + \epsilon_i \ (1)$$

čia $y_i$ yra D dyždio dimensijos priklausomų kintamųjų vektorius, o vektorius $\beta_0$ apima kiekvienam iš D aibės priklausomų kintamųjų priskirtus laisvuosius narius, $x_i$ – daugiamatės dimensijos požymių vektorius, priklausantis i-tam stebiniui [28]. $B$ yra regresijos koeficientų matrica, kurios kiekvienas elementas $B_{pd}$ nusako p-tojo požymio poveikį d-ajam stebiniui. Svarbiausia šio metodo prielaida yra tai, kad D dydžio atsitiktinių paklaidų vektorius $\epsilon_i$ yra laikomas nepriklausomai ir vienodai pasiskirsčiusiu visoje duomenų imtyje [28]. Klasikiniame daugialypės regresijos modelyje ši priklausomybė išreiškiama pasitelkiant daugiamatį normalųjį skirstinį tokiu būdu:

$$\epsilon_i \sim MVN(0, \sum) \ (2)$$

čia $MVN(0, \sum)$ yra daugiamatis normalusis (Gauso) skirstinys, kurio vidurkių vektorius lygus 0, o kovariacijų matrica yra teigiamai apibrėžta. Taikant daugialypę regresiją NT kainų modeliavime,

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html#sklearn.preprocessing.OneHotEncoder

dažnu atveju susiduriama su netenkinamomis prielaidomis. Viena iš priežasčių – didelis paaiškinamųjų požymių skaičius. Tokios problemos sprendimui dažniausiai taikomi reguliarizacijos metodai, naudojant šiuos metodus papildomai pritaikomas baudos dydis regresijos koeficientams, tokiu būdu padidinamas bendras modelio tikslumas. Šiame tyrime daugiau dėmesio skirta 3 pagrindiniams reguliarizuotos regresijos modeliams – *Lasso*, *Ridge* bei *ElasticNet*. Naudojant *Lasso* (angl. *Least Absolute Shrikage and Selection Operator*) metodą, į modelio tikslo funkcijos optimizavimo procesą įtraukiama *L1* normos baudos komponentė, tada $\beta$ koeficientų minimizavimo uždavinys sprendžiamas tokiu būdu [29]:

$$\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right\} \quad (3)$$

čia $y_i$ yra i-osios eilutės tikroji priklausomojo kintamojo reikšmė, $x_i^T\beta$ yra i-osios eilutės modelio prognozuota reikšmė, $\alpha$ yra reguliarizacijos koeficiento dydis, o $\sum_{j=1}^{p}|\beta_j|$ yra *L1* normos bauda, pagal kurią sumažinamos visų $\beta$ koeficientų reikšmės. Priešingai nei *Lasso*, naudojant *Ridge* metodą, koeficientų reikšmės nėra „nubaudžiamos“ iki nulio, įtraukiama *L2* normos baudos komponentė, tada minimizavimo uždavinys sprendžiamas tokiu būdu [29]:

$$\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \alpha\sum_{j=1}^{p}\beta_j^2\right\} (4)$$

čia $\alpha\sum_{j=1}^{p}\beta_j^2$ yra *L2* normos baudos komponentė. *ElasticNet* metodas apjungia pateiktus reguliarizavimo principus, tokiu būdu galima išnaudoti tiek *Lasso*, tiek *Ridge* metodų pranašumus:

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\right\} (5)$$

čia *lambda* yra bendras reguliarizacijos valdymo parametras. *ElasticNet* metodas yra naudingas, kai duomenyse egzistuoja didelis skaičius koreliuojančių požymių ir nežinoma, kuris reguliarizacijos metodas pasižymės didesniu efektyvumu [29]. Visiems modeliams svarbų vaidmenį atlieka hiperparamteras $\alpha$, kuris kontroliuoja reguliarizavimo stiprumą. Esant didesnei šio parametro reikšmei, regresijos koeficientai yra labiau nubaudžiami, ir atvirkščiai, labai maža $\alpha$ reikšmė leidžia modelio koeficientams laisviau prisitaikyti prie duomenų, tačiau tai gali padidinti „triukšmo“ įtaką modeliavimo rezultatams. Svarbu paminėti, kad šiame darbe reguliarizuotos regresijos modeliai buvo išbandyti atliekant ir svarbių paaiškinamųjų požymių atranką, tada hiperparametras $\alpha$ buvo parenkamas automatiškai, pasitelkiant *LassoCV* ir *RidgeCV* funkcijas iš *sklearn* bibliotekos.

Taip pat išbandyta keletas alternatyvių modelių, vienas iš jų – Bajeso teorema paremtas *Bayesian Ridge* modelis, kuriame regresijos koeficientai traktuojami kaip atsitiktiniai dydžiai, turintys iš anksto apibrėžtus tikimybinius skirstinius. Naudojant šį modelį laikomasi prielaidos, kad koeficientai turi sferinį daugiamačio normaliojo skirstinio priorą, kurio dispersiją apibrėžia hiperparametras $\lambda$. Tuo tarpu triukšmo dispersiją apibūdiną parametras $\alpha$, o abiejų hiperparmetrų skirstiniai modeliuojami naudojant *gamma* skirstinius [30]. Skirtingai nei klasikiniuose regresijos modeliuose, naudojant šį metodą, minimizuojama ne tik paklaidų kvadratų suma, bet ir įvertinami regresijos koeficientų pasiskirstymai.

Kitas naudingas metodas – stochastinio gradientinio nusileidimo (angl. *Stochastic Gradient Descent*) regresija. Tokio tipo modelis kiekvienoje optimizacijos iteracijoje naudoja atsitiktines stebinių imtis ir reguliariai atnaujina regresijos koeficientus pagal gradiento informaciją. Modelyje gali būti naudojamos skirtingos nuostolių funkcijos, pavyzdžiui, naudojant *Huber* nuostolių funkciją su *L2* baudos komponente, optimizuojama tikslo funkcija atrodytų taip:

$$\min_{\beta} \sum_{i=1}^{n} \mathcal{L}_{\delta}(y_i, x_i\beta) + \alpha\|\beta\|_2^2 \quad (6)$$

čia $\mathcal{L}_{\delta}$ – *Huber* funkcija, kuri elgiasi kaip kvadratinė funkcija, kai paklaidos yra mažos, ir kaip tiesinė, kai paklaidos yra didelės, tokiu būdu sumažinama išskirčių įtaka modeliavimo rezultatams. Šiame darbe naudotas *SGD* modelis buvo konfigūruotas su parametru *penalty='l2'*, tokiu būdu panaudojama *Ridge* tipo reguliarizacija. Taip pat išbandyta *adaptive* parinktis, kuri leidžia dinamiškai keisti mokymosi žingsnį.

Šiame darbe taip pat išbandyta atraminių vektorių regresija (angl. *Support Vector Regression*). Šio metodo esmė – surasti tokią funkciją, kuri prognozuotų priklausomojo kintamojo reikšmes taip, kad dauguma prognozės paklaidų patektų į iš anksto apibrėžtą tolerancijos ribą $\epsilon$. *SVR* pasižymi lankstumu – tokio tipo modelis, priklausomai nuo duomenų struktūros, gali naudoti įvairių tipų funkcijų branduolius [29]. Atsižvelgiant į duomenų struktūrą, pasirinktas radialinio pagrindo funkcijos branduolys (angl. *radial basis kernel*), kuris modeliui leido efektyviau įvertinti netiesinius ryšius tarp paaiškinamųjų požymių ir pasirinkto priklausomojo kintamojo [29]. Be šio parametro, taip pat konfigūruotos ir baudos koeficiento *C* ir branduolio funkcijos parametro *gamma* reikšmės.

Taip pat pritaikytas *K* artimiausių kaimynų metodas, kuris paremtas idėja, kad priklausomojo kintamojo reikšmes galima prognozuoti pagal jam artimiausių duomenų taškų (kaimynų) reikšmes. Kitaip tariant, prognozuojant tam tikro buto kainą, algoritmas ieško *K* artimiausių stebėjimų paaiškinamųjų požymių erdvėje ir skaičiuoja prognozę pagal šių stebėjimų kainų vidurkį (arba pagal svorinį vidurkį, kai „arčiau" esantiems kaimynams suteikiamas didesnis svoris) [31]. Naudojant K artimiausių kaimynų regresijos modelį buvo optimizuoti kaimynų skaičiaus, atstumo skaičiavimo metodo ir svorio strategijos parametrai.

Apibendrinant šį skyrių, toliau pateikiami visi paminėti modeliai ir jų pavadinimai *sklearn* bibliotekoje:

- tiesinė regresija (*LinearRegression*)
- *Lasso* regresija (*Lasso*)
- *Ridge* regresija (*Ridge*)
- *ElasticNet* regresija (*ElasticNet*)
- Bajeso teorema paremta *Ridge* regresija (*BayesianRidge*)
- stochastiniu gradiento nusileidimu paremta regresija su *Huber* nuostolio funkcija (*SGDRegressor*)
- atraminių vektorių mašinos SVR regresija (*SVR*)
- K artimiausių kaimynų regresija (*KNeighborsRegressor*)

## 2.4. Sprendimų medžių ansamblių regresijos modeliai

Sprendimų medžių ansambliu paremti metodai pastaraisiais metais tapo vienu iš patikimiausiu pasirinkimų sprendžiant regresijos ir klasifikavimo uždavinius, ypač kai nagrinėjami sudėtingų struktūrų duomenys, kaip dažnai pasitaiko NT rinkos atveju. Skirtingai nei su klasikiniais tiesinės regresijos modeliais, naudojant sprendimų medžių ansamblių regresiją prielaidos apie duomenų pasiskirstymus nėra tokios svarbios, taip pat efektyviai apdorojami tiek skaitiniai, tiek kategoriniai požymiai [32]. Šie veikimo principai yra itin naudingi konstruojant patikimą vertinimo modelį, kuris buto kainos prognozes skaičiuotų ne tik pagal atskirus požymius, bet ir pagal jų tarpusavio sąveikas.

Pats sprendimų medis pasižymi paprasta struktūra, šis modelis dažnai kenčia nuo per didelio prisitaikymo treniravimo duomenims (angl. *overfitting*). Dėl šios priežasties praktikoje taikomi minėti medžių ansamblių metodai. Pagrindinė idėja – apjungti daugybę silpnų prognozavimo modelių į vieną bendrą sistemą, tokiu būdu gauti tikslesnius ir stabilesnius rezultatus [32]. Pagal veikimo principus šie metodai gali būti suskirstyti į dvi pagrindines kategorijas. Pirmoji – tai metodai, paremti visų sprendimų medžių vidurkiu, naudojant šiuos metodus, kiekvieno medžio skaičiuojama prognozė yra nepriklausoma, o galutinė reikšmė apskaičiuojama kaip visų prognozių vidurkis. Tokie metodai padeda sumažinti kuriamų modelių variaciją ir užtikrina sąlyginai stabilias prognozes, ypač kai kiekvienas medis treniruojamas su skirtingais duomenų poaibiais (angl. *boostrap sampling*). Ši strategija plačiai žinoma kaip *bagging* [32]. Tačiau vis tiek egzistuoja vienas trūkumas – didelės tarpusavio koreliacijos tarp medžių. Šiai problemai išspręsti pasitelkiamas atsitiktinių miškų sumaišymo principas, kai kiekviename medyje atsitiktinai atrenkamas požymių pogrupis.

Antroji kategorija susideda iš pastiprinimo (angl. *boosting*) metodų, pagal šiuos metodus modeliai konstruojami nuosekliai, tai reiškia, jog kiekvienas naujas modelio vienetas (medis) treniruojamas su tikslu pataisyti prognozes, atsižvelgiant į ankstesnių medžių prognozavimo rezultatus [32]. Šis principas leidžia sukurti efektyvesnius modelius net iš labai paprastų duomenų rinkinių. Abi minėtos metodų kategorijos – tiek *bagging*, tiek *boosting*, yra laikomos kaip pagrindinės, pagal kurias kuriama daugybe šiuolaikinių NT objektų vertinimo modelių. Toliau pateikiami tyrime naudoti sprendimų medžių ansamblių regresijos modeliai ir jų pavadinimai *sklearn* bibliotekoje:

- atsitiktinio miško regresija (*RandomForestRegressor*)
- gradientinio pastiprinimo regresija (*GradientBoostingRegressor*)
- „Bagging“ metodu paremta regresija (*BaggingRegressor*)
- papildomų medžių regresija (*ExtraTreesRegressor*)
- „Cat“ pastiprinimo regresija iš *catboost* bibliotekos (*CatBoostRegressor*)
- ekstremalaus gradientinio pastiprinimo regresija (*XGBRegressor*)
- švelnios gradientinio pastiprinimo mašinos regresija (*LGBMRegressor*)
- histograma paremto gradientinio pastiprinimo regresija (*HistGradientBoostingRegressor*)
- „Ada“ pastiprinimo regresija (*AdaBoostRegressor*)

## 2.5. Modelių vertinimo kriterijai

Vertinant sukurtų regresijos modelių tikslumą tyrime buvo taikomi du pagrindiniai kokybės rodikliai – vidutinė kvadratinės šaknies paklaida (*RMSE*) ir vidutinė absoliuti procentinė paklaida (*MAPE*), šie rodikliai išreiškiami tokiu būdu [33]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

čia $y_i$ yra tikroji stebėjimo reikšmė, $\hat{y}_i$ yra modelio prognozė, o $n$ – stebėjimų skaičius. Naudojant šiuos du rodiklius, modeliai buvo vertinti taikant penkių kartų kryžminio patikrinimo procedūrą. Tai reiškia, jog visa duomenų imtis buvo atsitiktinai padalinta į penkias lygias dalis, kiekvieną kartą viena iš dalių naudojama kaip testavimo rinkinys, o likusios keturios skirtos modelio treniravimui. Galutinis modelio tikslumas įvertintas naudojant kryžminio patikrinimo metu gautus kokybės kriterijų vidurkius. Tokia procedūra leido gauti patikimesnius rezultatus bei objektyviau palyginti skirtingų modelių prognozavimo kokybę.

### 2.6. Pagrindinių komponenčių analizė

Nagrinėjant didelio masto duomenų rinkinius, kuriuose egzistuoja šimtai požymių, dažnai atsiranda poreikis sumažinti duomenų dimensijas, neprarandant esminės informacijos. Tokiu atveju vienas plačiausiai taikomų metodų yra pagrindinių komponenčių analizė (angl. *Principal Component Analysis*). Naudojant *PCA*, bandoma rasti tokį požymių transformacijos vektorių rinkinį *W*, kuris užtikrintų maksimalią transformuotų duomenų dispersiją [34]:

$$Z = XW \quad (9)$$

čia *Z* – sumažintų dimensijų duomenys, o *X* – pradinių požymių matrica. Vektoriai *W* randami atliekant duomenų kovariacijos matricos spektrinę dekompoziciją [34].

Praktinėje šio darbo dalyje *PCA* įrankis buvo taikytas kaip papildomas struktūrinių parametrų elementas, integruotas į modeliavimo etapą, naudojant grandinės (angl. *Pipeline*) konstrukciją. Programos kode tai įgyvendinta nurodant parametrą *n_component=0.95*, tai reiškia, kad modelis atrinks tiek pagrindinių komponenčių, kiek reikia, kad būtų išlaikyta 95% visos duomenų dispersijos.

### 2.7. Požymių transformacijos

Siekiant dar kartą pagerinti regresijos modelių tikslumą, šiame tyrime buvo išbandyta keletas alternatyvių požymių inžinerijos priemonių – papildomų polinominių požymių kūrimo funkcija *PolynomialFeatures* ir tolydžiųjų požymių skirstinių transformacijos funkcija *PowerTransform*. Pirmasis įrankis leido iš pradinių duomenų rinkinių sugeneruoti papildomus sąveikų bei aukštesnio laipsnio požymius. Pavyzdžiui, jei pradiniame rinkinyje yra požymiai $x_1$ ir $x_2$, tai panaudojant *PolynomialFeatures* sukūriamos papildomos sąveikos, tokios kaip $x_1 * x_2, x_1^2, x_2^2$ ir t. t. [35]. Taikant šią funkciją, duomenų rinkiniai buvo papildyti skaitinių požymių antro laipsnio polinominėmis sąveikomis. *PowerTransform* įrankis leido modifikuoti tolydžiuosius požymius, kurių pasiskirstymas reikšmingai skyrėsi nuo Gauso skirstinio [36]. Greta šių priemonių taip pat taikytos ir paprastosios skaitinių požymių skalių transformacijos – standartizavimo, normalizavimo, ir išskirtims atspari *RobustScaler*. Panašiai kaip ir su *PCA* įrankiu, visos požymių transformacijos išbandytos naudojant modeliavimo grandinės konstrukciją, kai kiekvienam turimam duomenų rinkiniui, naudojant 5 kartų kryžminį patikrinimą, testuojamos visos turimos kombinacijos, tokiu būdu identifikuojamas

konkretaus duomenų rinkinio požymių transformacijų sprendimas, užtikrinantis geriausius prognozavimo rezultatus [26].

## 2.8. Statistiškai reikšmingų požymių atrankos metodai

Požymių atranka – dar vienas svarbus etapas sprendžiant regresijos uždavinius, šio proceso metu sumažinamas modelių kompleksiškumas ir neigiama „perteklinių“ požymių įtaka prognozavimo tikslumui. Šiame darbe taikyti du pagrindiniai atrankos metodai – reguliarizuotų regresijos modelių analizė bei medžių ansambliais grįsta atranka. Pirmasis metodas paremtas *Lasso* ir *Ridge* regresijos algoritmais ir jų baudos koeficientais, kurie sumažina statistiškai nereikšmingų požymių įtaką arba ją visiškai pašalina [26], [37]. Požymiai buvo atrinkti remiantis jų koeficientų reikšmėmis – į galutinius rinkinus patekdavo tie požymiai, kurių koeficientų absoliutinė vertė viršijo iš anksto nustatyta reikšmę (angl. *threshold*).

Antrasis tyrime taikytas atrankos metodas paremtas *Boruta* algoritmu, kuris naudoja standartinį atsitiktinio miško modelį. Metodo esmė – palyginti tikrųjų paaiškinamųjų požymių svarbumą su „šešėliniais“ (angl. *shadow features*) požymiais, kurie gaunami permutavus pradinius duomenis. Jei tikrasis požymis statistiškai reikšmingai pranoksta bet kurį šešėlinį variantą, jis yra identifikuojamas kaip „vertingas“ prognozuojant priklausomąjį kintamąjį [26], [38]. Atsižvelgiant į tai, jog kiekvieno modelio prognozių tikslumas gali būti pagerintas naudojant skirtingus kiekius požymių, modeliavimo etape kiekvienam duomenų rinkiniui išbandyta tiek reguliarizuotų regresijų, tiek *Boruta* algoritmo požymių atranka. Tai reiškia, jog kiekvienam pradiniam požymių rinkiniui pirmiausia sukurtas papildomas, sudarytas tik iš *Lasso* ir *Ridge* modelių rezultatų, tuomet kitas, sudarytas iš *Boruta* algoritmo rezultatų, o vėliau gautas paskutinis, mišrusis rinkinys, apjungiantis skirtingų metodų pagalba identifikuotus reikšmingus požymius. Modeliavimo etape visi rezultatai yra lyginami su originaliais (nenaudojant *PolynomialFeatures* funkcijos) požymių rinkiniais, tokiu būdu įvertinant, ar atrankos procedūra iš tiesų padeda pagerinti modeliavimo rezultatus.

## 2.9. Modelių parametrų optimizavimas

Ankstesniuose skyriuose išvardintų modelių struktūrinių parametrų optimizavimas buvo įgyvendintas naudojant *GridSearchCV* įrankį, kuris leidžia išbandyti visas kombinacijas ir pritaikyti k-kryžminį patikrinimą. Tačiau reikėtų atkreipti dėmesį, kad naudojant šį metodą modelių vidinių parametrų (hiperparametrų) optimizavimui, susiduriama su esmine problema, kai egzistuojančių kombinacijų skaičius gali būti begalinis. Tokiu atveju *GridSearchCV* tampa neefektyvus – šio įrankio paieškos algoritmas paremtas visų kombinacijų išbandymu, naudojant jį modeliuose su keliais tolydžiais hiperparametrais, niekada nebus pasiektos optimalios reikšmės. Taip pat egzistuoja alternatyvi priemonė – *RandomizedSearchCV*, tačiau ir ji nėra tinkama hiperparametrų optimizavimui, nes neatsižvelgia į ankstesnių bandymų rezultatus.

Dėl šių priežasčių šiame darbe buvo ieškota efektyvesnių optimizavimo įrankių. Internete daugiausia informacijos galima rasti apie 3 skirtingas Python kalbos bibliotekas (*Raytune*, *Optuna* ir *HyperOpt*), visos jos pasižymi efektyvios paieškos algoritmų įvairove. Šiame darbe pasirinkta naudoti *Optuna*, nes šis įrankis pasižymi geru suderinamumu su *sklearn* bibliotekos modeliais [26]. Vietoje speliojimu paremtos paieškos, naudojant *Optuna*, algoritmas gali prisitaikyti prie optimizacijos krypties pagal ankstesnių bandymų rezultatus [39]. Tai įgyvendinama panaudojant *Tree-structured Parzen*

*Estimator*[10] (*TPE*) algoritmą. Pirmiausia parenkama 10 atsitiktinių parametrų kombinacijų, o nuo vienuoliktosios iteracijos pradedama remtis ankstesniais rezultatais – naudojant *TPE* algoritmą įvertinama, kurios hiperparametrų sritys turės daugiausia įtakos prognozių tikslumui [39]. Tokiu būdų bendra paieškos erdvė tiriama kryptingai, taip pat sumažinamas konvergavimui reikalingų bandymų skaičius. Praktinėje tyrimo dalyje kiekvienas regresijos modelis buvo optimizuotas pagal *objective* funkciją, kuri įvertina kiekvieno išbandyto modelio prognozavimo paklaidą (šioje dalyje pasirinkta naudoti *MAPE*). Optimizavimo metu kiekvienam modeliui apibrėžtas parametrų paieškos laukas, tada kiekvienai kombinacijai buvo skaičiuojama *MAPE* reikšmė, taip pat fiksuotas kiekvieno modelio optimizavimui reikalingas bendras skaičiavimo laikas. Optimizuojant modelių hiperparametrus nustatytas maksimalus leidžiamų bandymų skaičius – 100 hiperparametrų kombinacijų, arba alternatyvus 30 minučių skaičiavimo laiko apribojimas, priklausomai nuo to, kuris kriterijus buvo išpildytas greičiau.

[10] https://optuna.readthedocs.io/en/stable/reference/samplers/generated/optuna.samplers.TPESampler.html

## 3. Mokslinis tyrimas

Mokslinis tyrimas gali būti išskaidytas į kelis pagrindinius etapus:

1. Pirmiausia atlikta žvalgomoji duomenų analizė, kurios metu vizualizuoti turimi duomenys, pašalinti nereikalingi požymiai, duomenų rinkiniai praturtinti naujais požymiais, užpildytos trūkstamos reikšmės;
2. Kitame etape atliktas struktūrinių parametrų optimizavimas – parinktas geriausias priklausomas kintamasis, įvertinta kategorinių požymių, pagrindinių komponenčių analizės bei skaitinių požymių skalių transformacijų nauda. Atradus geriausius struktūrinių parametrų sprendimus, patikrinta, ar skaitinių požymių sąveikos bei atitinkamos skirstinių transformacijos padidina prognozių tikslumą;
3. Remiantis skirtingais metodais, atlikta statistiškai reikšmingų paaiškinamųjų požymių atranka;
4. Naudojant skirtingus požymių rinkinius, atliktas išsamus modelių hiperparametrų optimizavimas;
5. Identifikuoti didžiausiu tikslumu pasižymintys modeliai, taip pat nustatyta, kurie paaiškinamieji požymiai yra svarbiausi modeliavimo etape;
6. Naudojant 5 geriausius modelius, atliktas butų pardavimo ir nuomos kainų modeliavimas pagal atskirus duomenų rinkinių segmentus.

### 3.1. Žvalgomoji duomenų analizė

Kiekvienas pradinis duomenų rinkinys turėjo 26 paaiškinamuosius požymius, surinkti duomenys atrodė taip:

- Vilniaus butų pardavimo duomenys (2527 stebėjimai);
- Vilniaus butų nuomos duomenys (1562 stebėjimai);
- Kauno butų pardavimo duomenys (680 stebėjimų);
- Kauno butų nuomos duomenys (555 stebėjimai).

Užbaigus duomenų rinkimo procedūrą, atliktas pirminis įsivertinimas – siekta suprasti bendrą struktūrą ir iš karto identifikuoti galimas problemas. Pirmiausia atliktos pradinės duomenų failų modifikavimo procedūros:

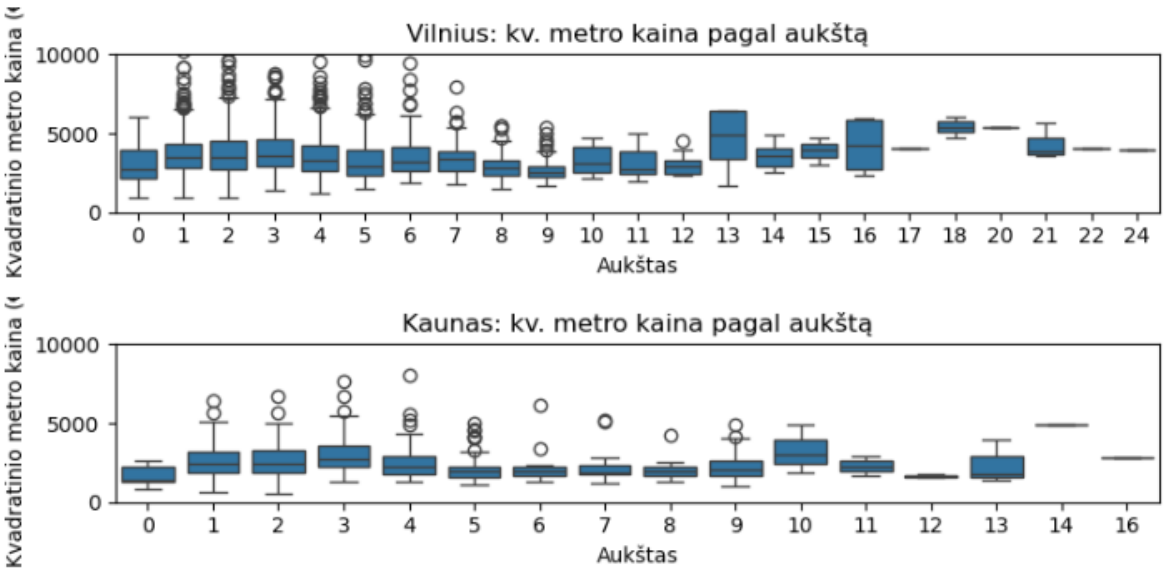
- sukurtas požymis „kv_kaina“, kuris buvo naudotas kaip alternatyvus priklausomojo kintamojo pasirinkimas;
- visuose duomenų rinkiniuose požymis „statybos_metai“ buvo konvertuotas į požymį „pastato_amzius“ ($2025 - \text{„}statybos\_metai\text{"}$);
- visi atstumai suvienodinti pagal kilometrų skalę;
- sukurtas požymis „plotas_per_kamb“.

Sekančiame etape kategoriniai požymiai buvo transformuoti į modeliams tinkamą formatą. Kadangi požymis „sildymo_tipas“ gali turėti daug skirtingų kombinacijų, jis buvo perkeltas į daugiamatį binarinį formatą, taikant „one-hot“ kodavimą. Atnaujintame duomenų rinkinyje kiekvienas stebėjimas pažymėtas loginėmis reikšmėmis, nurodant, ar konkretus butas turi tam tikrą šildymo tipą. Toliau vizualizuotos duomenų rinkinių trūkstamos reikšmės, tai leido greitai identifikuoti duomenų „pilnumo“ situaciją. Pastebėta, kad beveik visi duomenų rinkinių stulpeliai pasižymėjo mažu arba nuliniu trūkstamų reikšmių kiekiu – tai lėme jau duomenų surinkimo etape priimtas sprendimas rinkti tik svarbiausius, kiekviename skelbime pasitaikančius elementus. Vienintelės išimtys yra tik keli papildomi požymiai, pvz., oro taršos „no2“ ir „kd10“ rodikliai ir požymis „langu_orientacija“, kurio

trūkstamų reikšmių dalis viršijo 50% visuose duomenų rinkiniuose (žr. 5 pav.). Galima teigti, kad nėra kito būdo kaip tik atsisakyti šio požymio. Kita vertus, oro taršos rodikliai gali būti naudingi modeliavimo etape, todėl šių požymių trūkstamos reikšmės buvo užpildytos, pasitelkiant buto koordinates ir *K* artimiausių kaimynų metodą iš *sklearn* bibliotekos (parametras k=3).

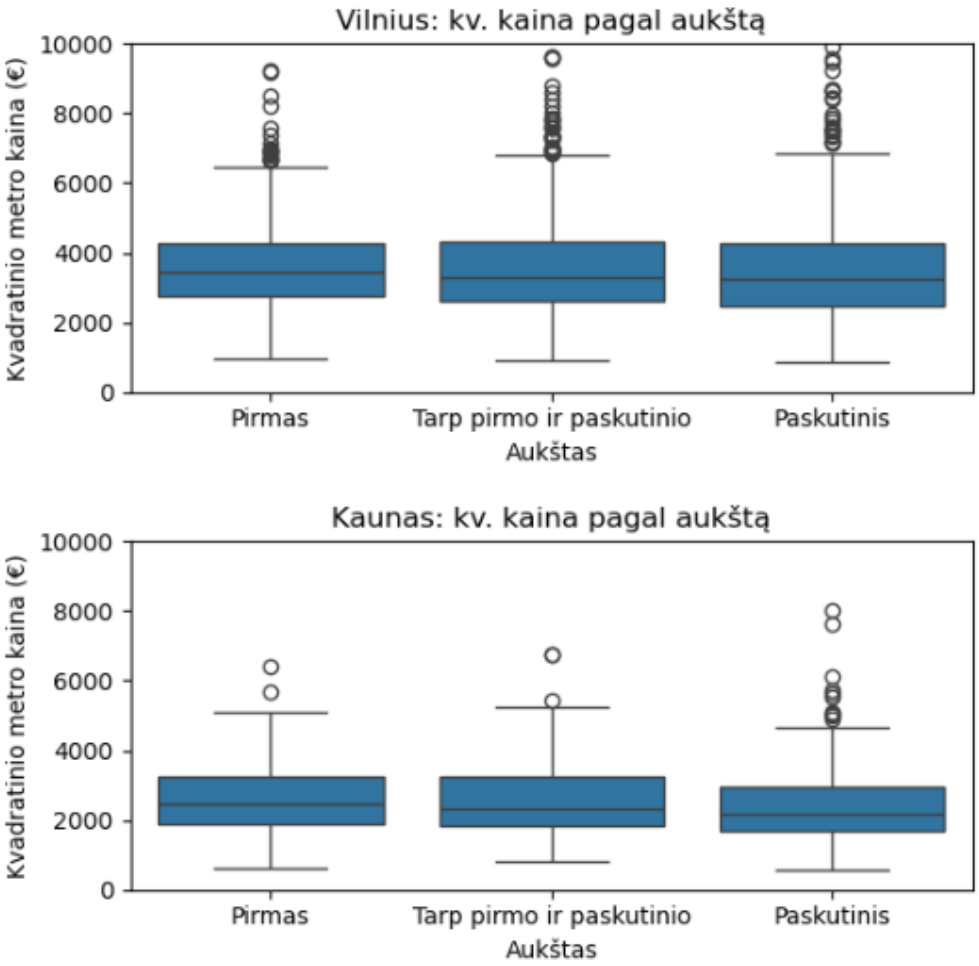

**5 pav.** Butų pardavimo duomenų rinkinių paaiškinamųjų požymių trūkstamos reikšmės

Perėjus prie papildomų požymių kūrimo, pirmiausia įvertinta, ar buto aukšto požymiai turi ryšį su prognozuojamu kintamuoju (šiuo atveju kvadratinio metro kaina). Pirmiausia vizualizuotas kvadratinio metro kainos pasiskirstymas atskirai kiekviename aukšte (žr. 6 pav.). Pastebėta, kad Vilniaus butų pardavimo duomenų rinkinyje kvadratinio metro kainos mediana pagal aukštą išlieka stabili, o didėjant buto aukštui, kyla pagal tendenciją, kai aukščiau esantis butas gali pasižymėti naujesniu įrengimu ir turėti geresnį vaizdą pro langus. Įdomu ir tai, kad Kauno atveju pastebima priešinga tendencija – ties 5 ir 11 aukštais užfiksuotas akivaizdus kvadratinio metro kainos sumažėjimas. Šiam reiškiniui daug įtakos turi butai, esantys senuose monolitiniuose pastatuose (ypatingai Eigulių, Šilainių ir Dainavos rajonuose).



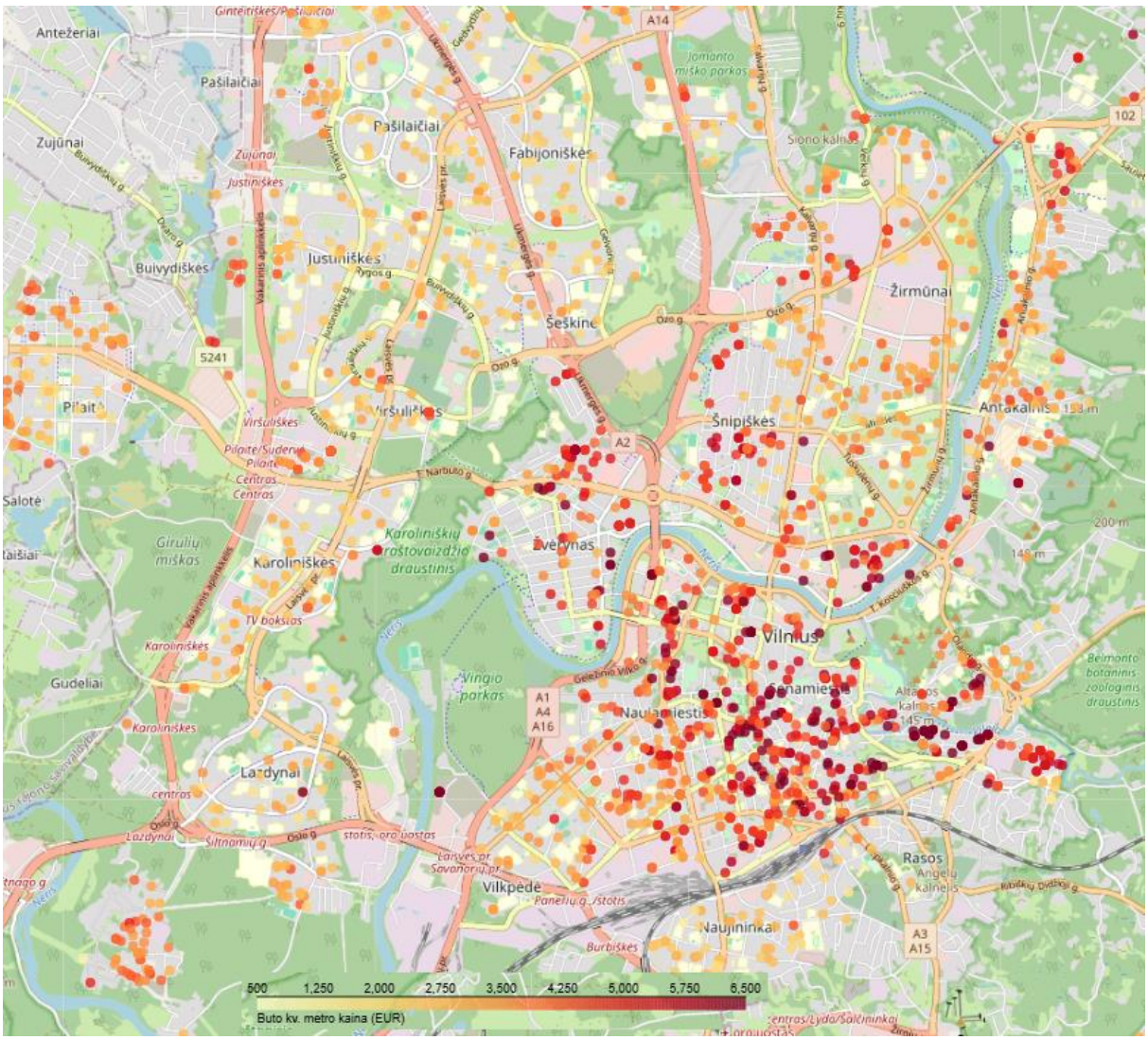

**6 pav.** Butų kv. metro pardavimo kainos pasiskirstymai pagal aukštą

Toliau bandyta pavaizduoti butų kvadratinio metro kainą pagal pirmą, paskutinį ir vidurinį aukštus. Remiantis pateiktu grafiku (žr. 7 pav.), buvo pastebėta, kad tiek Vilniaus, tiek Kauno butų pardavimo duomenų rinkiniuose paskutinio aukšto butai pasižymi žemesne kvadratinio metro kainos mediana, ši tendencija ryškiau pastebima Kauno butų pardavimo duomenyse. Viena iš galimų šio reiškinio priežasčių – paskutinio aukšto butai dažnai pasižymi prastesnėmis mikroklimato sąlygomis. Dėl šios priežasties paskutiniame aukšte esančių butų paklausa yra mažesnė, o tai turi neigiamą poveikį tokių butų pardavimo kainai. Atsižvelgiant į šiuos pastebėjimus, duomenų rinkiniams papildomai sukurtas binarinis požymis „ar_paskutinis“, nurodantis, ar konkretus butas yra paskutiniame aukšte.



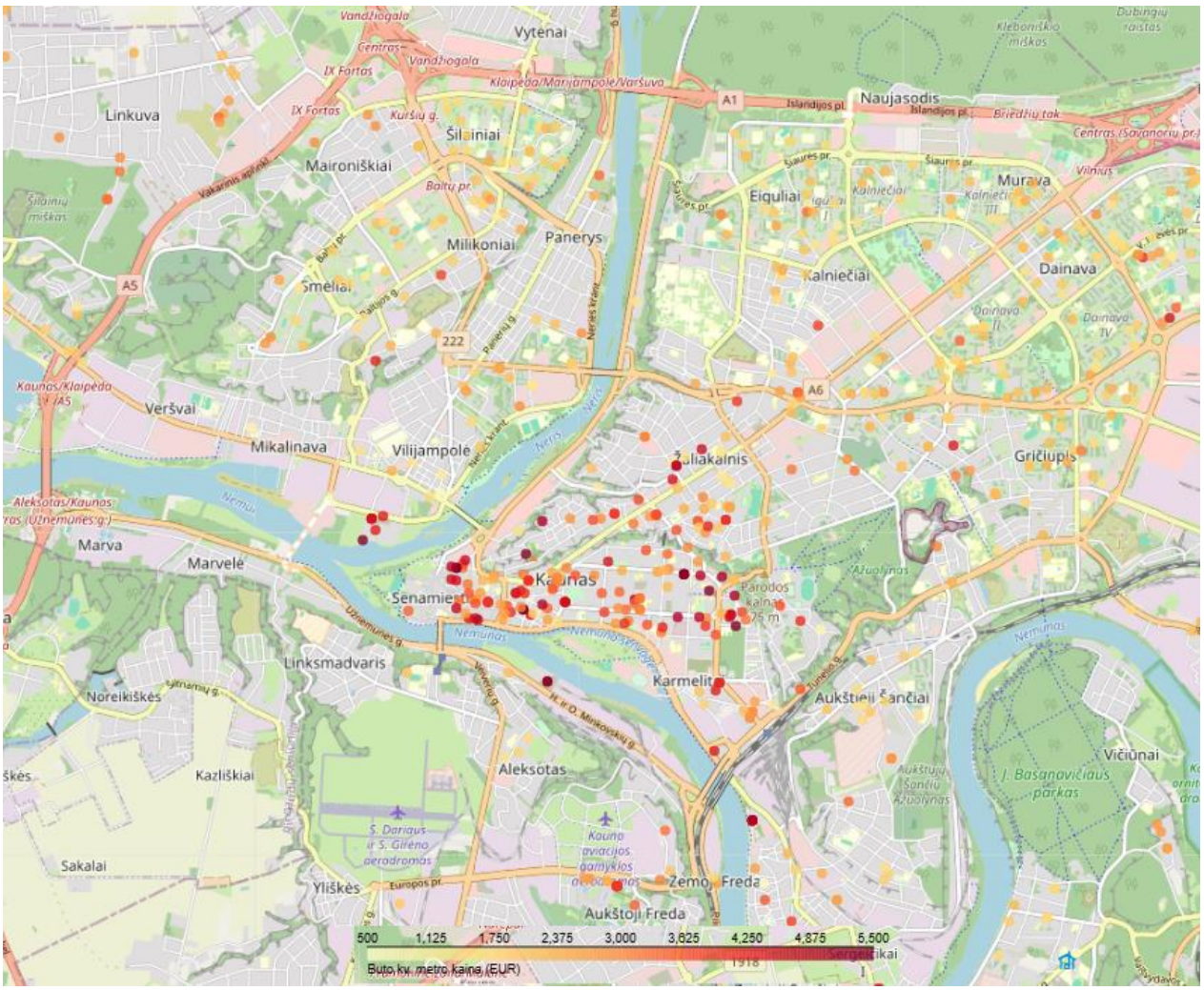

**7 pav.** Butų kv. metro pardavimo kainos pasiskirstymai pagal pirmą, paskutinį ir vidurinį aukštus

Sekančiame etape verta panagrinėti, koks yra butų pardavimo kainų ryšis ne tik struktūrinių, bet ir lokacinių požymių atžvilgiu. Nors surinktuose duomenyse jau yra keletas požymių apie atstumus iki reikšmingų objektų (viešojo transporto stotelių, ugdymo įstaigų, automobiliu važiuojant iki geležinkelio stoties ir pan.), vien šių požymių gali neužtekti, kai norime atskleisti buto pardavimo kainos ir lokacijos ryšį. Siekiant išplėsti analizę, bandyta sukurti naują požymį – euklido atstumą, matuojantį kiekvieno objekto nuotolį nuo apibrėžtų miesto centro koordinačių. Siekiant tiksliau apibrėžti, kur galėtų egzistuoti „centriniai“ taškai, buvo sukurti butų kvadratinio metro pardavimo kainų pasiskirstymų žemėlapiai (žr. 8 ir 9 pav.).



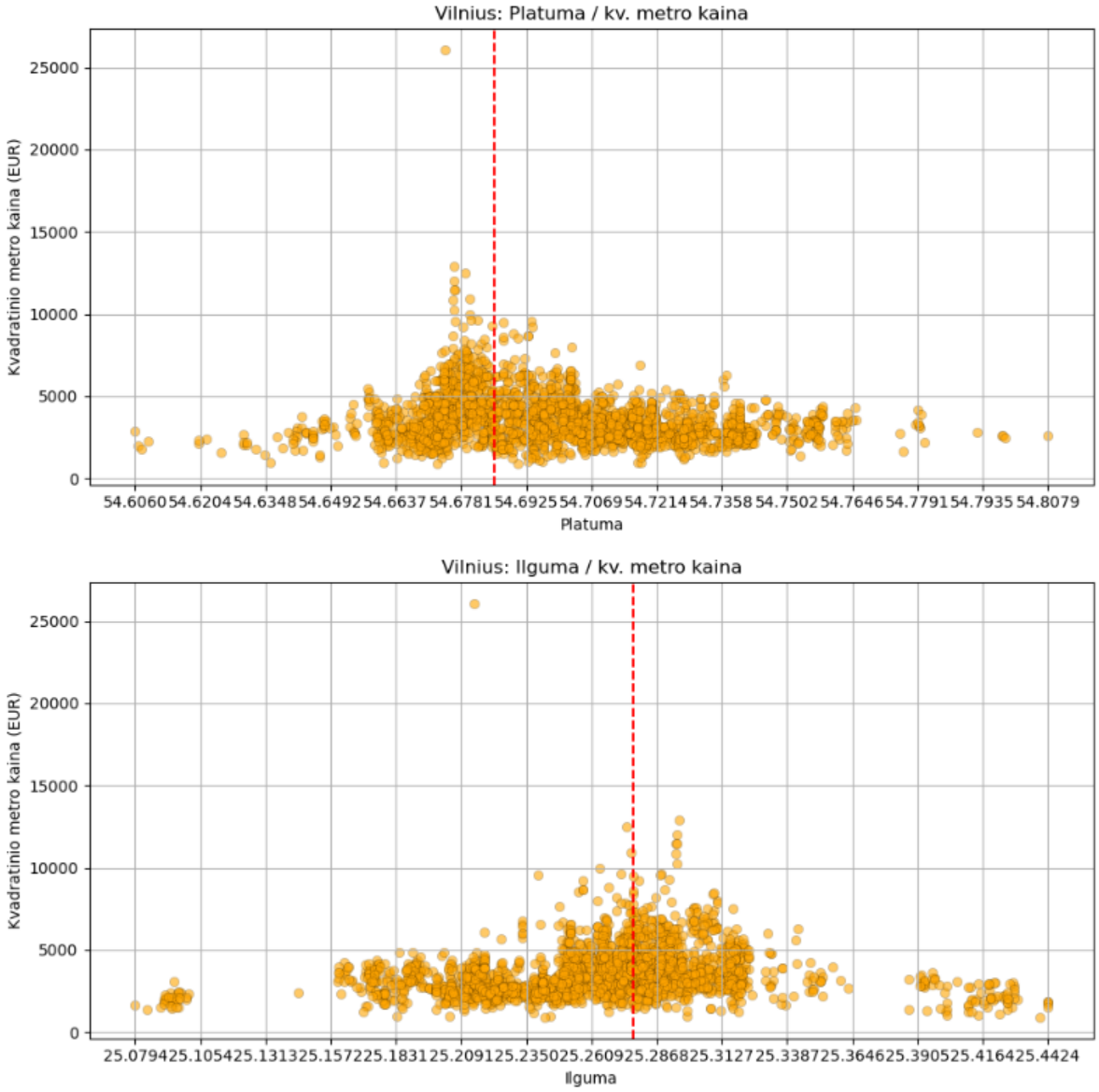

**8 pav.** Vilniaus miesto butų pardavimo kainų žemėlapis

Pagal 8 pav. rezultatus galima pastebėti, kad brangiausi objektai (pagal kvadratinio metro pardavimo kainą) yra išsidėstę aplink Senamiesčio, Žvėryno ir Šnipiškių rajonus, taip pat yra nemažai išskirčių šiauriniuose miesto rajonuose. Tokiu pačiu principu pavaizduotas ir Kauno miesto butų kvadratinio metro pardavimo kainų žemėlapis (žr. 9 pav.)



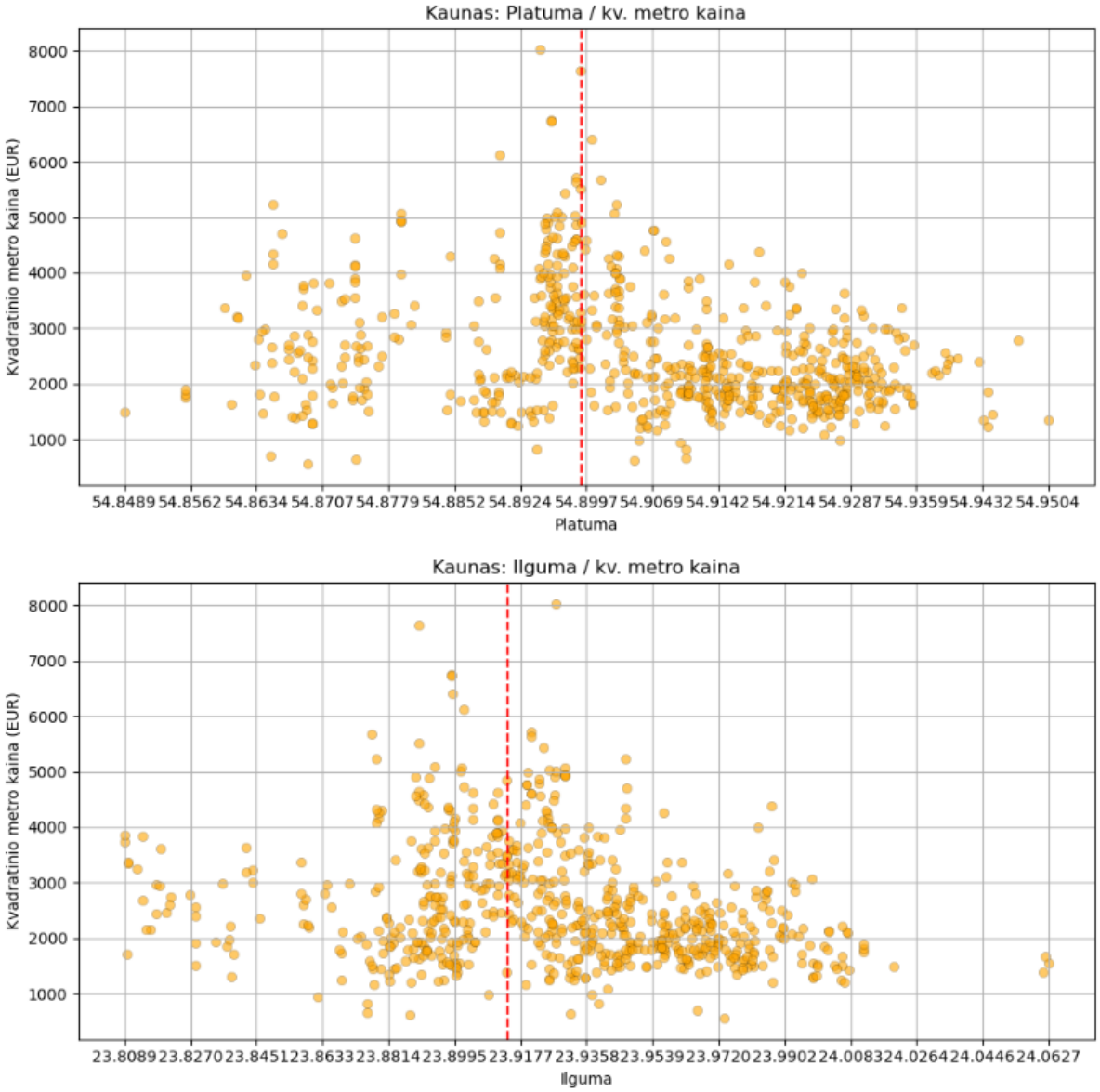

**9 pav.** Kauno miesto butų pardavimo kainų žemėlapis

Pagal 9 pav. rezultatus galima pastebėti, kad Kauno butų pardavimo duomenų atžvilgiu situacija yra kitokia, matosi aiškus ryšys tarp kvadratinio metro pardavimo kainos ir atstumo iki miesto centro. Taip pat pastebimi ir aiškiau atsiskyrę klasteriai – brangiausi objektai yra išsidėstę Senamiesčio rajone ir aplink Kauno soborą. Vis dėlto, iš 8 ir 9 pav. pateiktos informacijos nėra lengva susidaryti aiškią nuomonę, kas galėtų būti „centriniai“ taškai pardavimo kainos atžvilgiu, todėl toliau bandyta atvaizduoti kvadratinio metro pardavimo kainos ir platumos bei ilgumos grafikus, pradėta nuo Vilniaus butų pardavimo duomenų analizės (žr. 10 pav).




**10 pav.** Vilniaus butų kv. metro pardavimo kainos pasiskirstymai pagal platumos ir ilgumos reikšmes

Pagal 10 pav. pateiktą vizualizaciją „centrinio“ taško identifikavimo uždaviniui galima išskirti 2 sprendimo metodus – ekspertinės įžvalgos arba matematiniai algoritmai. Galima matyti, kad „centrinis“ taškas Vilniaus butų kvadratinio metro pardavimo kainos atžvilgiu turėtų būti tarp 54,6781 ir 54,6925 pagal platumą, ir tarp 25,2609 ir 25,2868 pagal ilgumą, tačiau tiksli reikšmė vis tiek nėra aiški. Dėl šios priežasties toliau išbandytas branduolio tankio įvertinimas (angl. *Kernel Density Estimation*) iš *sklearn.neighbors*. Pritaikius *KDE* metodą tiek platumos, tiek ilgumos duomenims, eksperimentuota su skirtingomis parametrų reikšmėmis. Geriausiai vizualiai atrodantis rezultatas gautas naudojant *bandwidth=0,2* ir *grid_resolution=100* parametrų reikšmes. Raudonos punktyrinės linijos žymi *KDE* metodo pagalba gautus taškus. Tokia pati procedūra atlikta naudojant ir Kauno butų pardavimo duomenis (žr. 11 pav.).



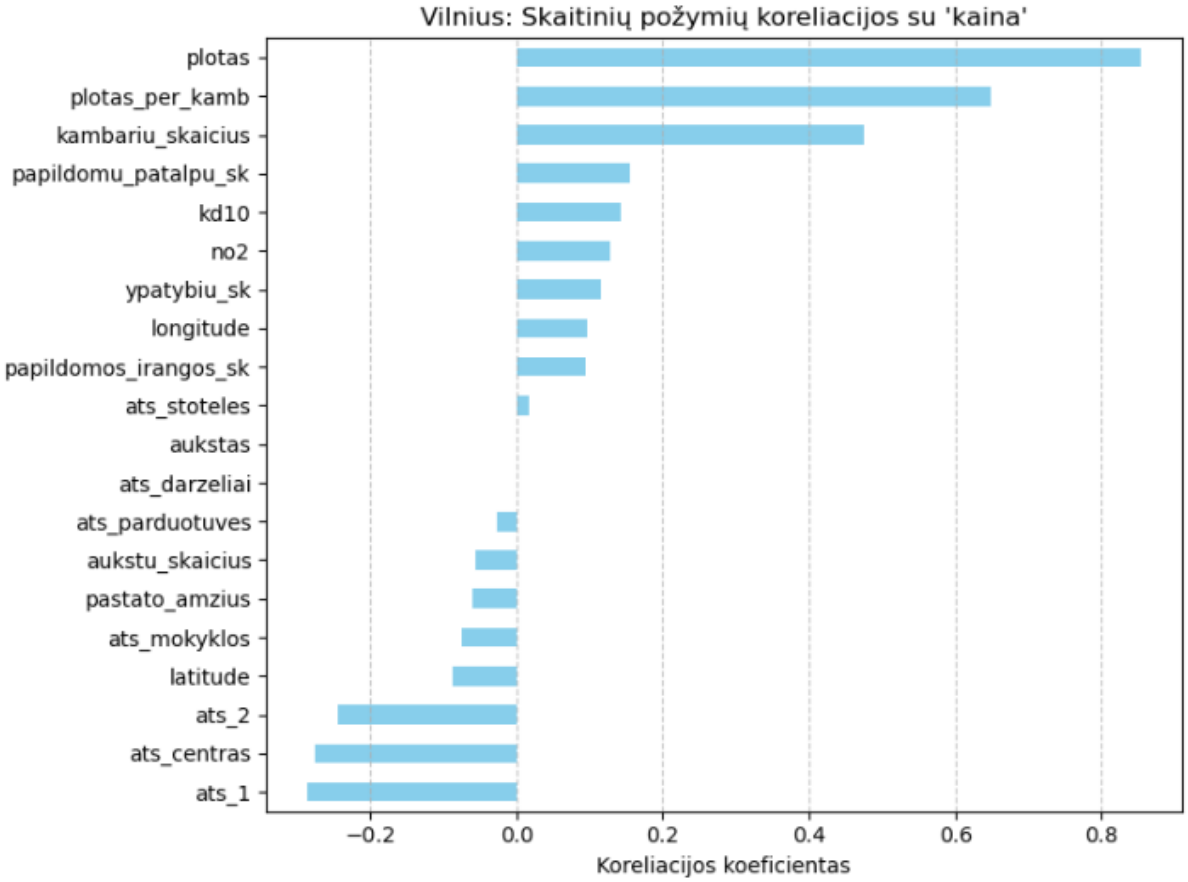

**11 pav.** Kauno butų kv. metro pardavimo kainos pasiskirstymai pagal platumos ir ilgumos reikšmes

Pagal pateiktą vizualizaciją galima pastebėti, kad Kauno atveju butų kvadratinio metro pardavimo kainos pasiskirstymas pagal platumos ir ilgumos požymius yra mažiau koncentruotas, lyginant su Vilniaus butų pardavimo duomenimis. Nepaisant šių skirtumų, buvo priimtas sprendimas dar kartą remtis *KDE* metodo rezultatais ir nustatyti preliminarius „centrinius“ taškus. Abiejų miestų taškai buvo dar kartą įvertinti, naudojant atnaujintus žemėlapius (žr. 1 ir 2 priedus). Vilniaus atveju gautas taškas yra prie Vilniaus arkikatedros, šis rezultatas nestebino, atsižvelgiant į tai, kad prabangiausi butai yra išsidėstę aplink Gedimino prospektą ir Senamiesčio rajoną. Kauno atveju *KDE* metodo pagalba gautas taškas yra netoli Vienybės aikštės. Koordinatės, kurios buvo naudotos skaičiuojant Euklido atstumus ir kuriant naują „ats_centras“ požymį, buvo tokios:

- Vilnius: platuma 54,6855, ilguma 25,2773;
- Kaunas: platuma 54,8991, ilguma 23,9139.

Siekiant geriau suprasti, kurie skaitiniai požymiai pasižymi stipriausiu ryšiu su buto pardavimo kaina, atlikta koreliacinė analizė. Šis žingsnis yra svarbus ruošiantis modeliavimo etapui, nes koreliacinės

analizės rezultatai yra naudingi ne tik sprendžiant multikolinearumo problemą, bet ir nustatant, kurių skaitinių kintamųjų sąveikas vertėtų išbandyti konstruojant skirtingus požymių rinkinius. 12 pav. pateikti Vilniaus butų pardavimo duomenų rinkinio koreliacinės analizės rezultatai.

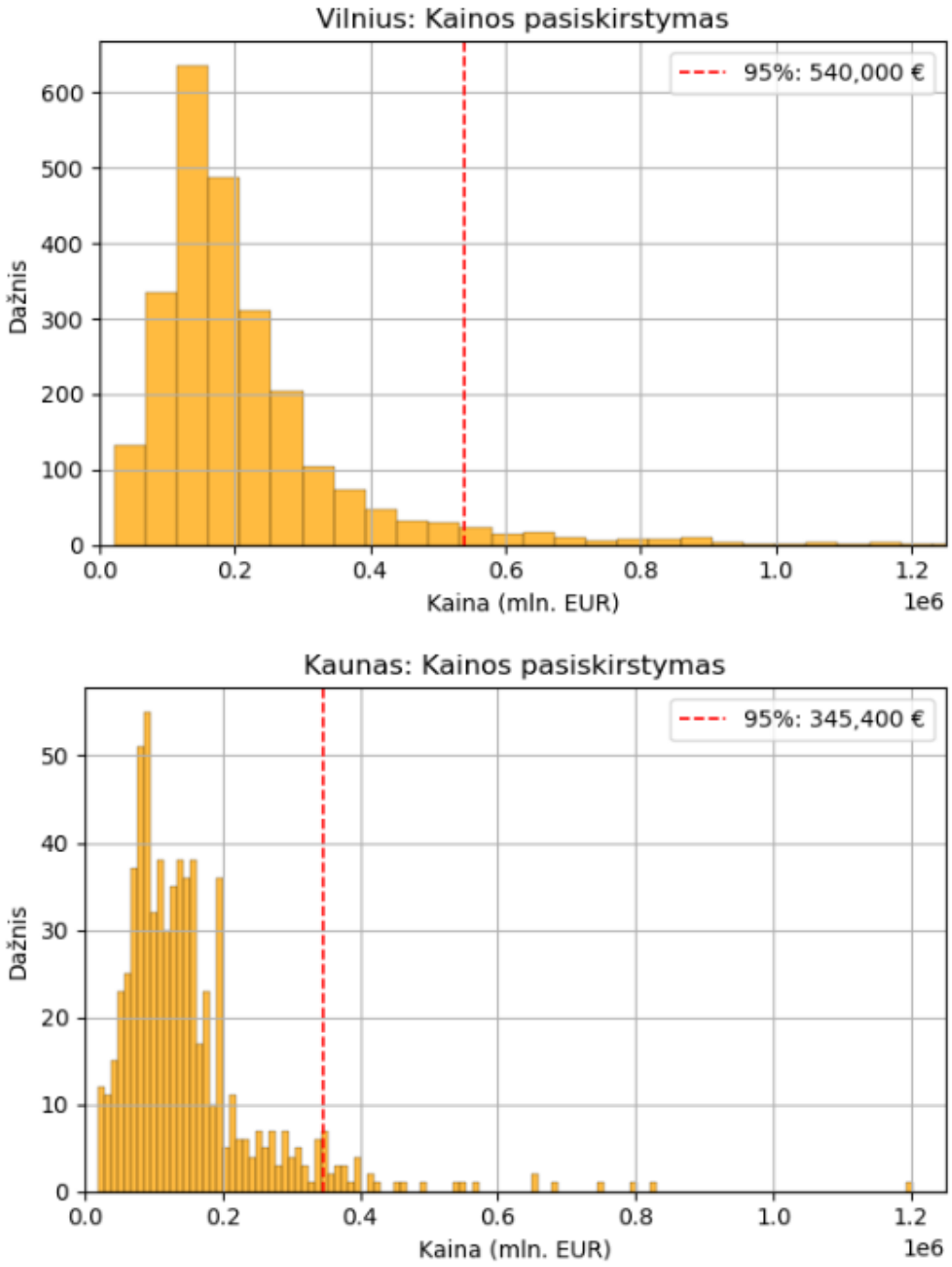



**12 pav.** Vilniaus butų pardavimo duomenų rinkinio skaitinių požymių ir buto kainos tarpusavio koreliacijos

Pagal 12 pav. rezultatus galima matyti, kad Vilniaus atveju egzistuoja stiprios koreliacijos tarp buto pardavimo kainos ir tokių požymių kaip bendras plotas, vieno kambario plotas bei kambarių skaičius. Šie rezultatai nestebina, nes stipriausiai koreliuojantys kintamieji visada yra esminiai kainą apibūdinantys požymiai. Be to, pastebėta, kad kai kurie požymiai neigiamai koreliuoja su buto pardavimo kaina, pvz., atstumas automobiliu važiuojant iki Vilniaus arkikatedros („ats_1“) arba iki geležinkelio stoties („ats_2“), taip pat naujai sukurtas Euklido atstumo požymis „ats_centras“. Kauno butų pardavimo duomenų rinkinio koreliacinės analizės rezultatuose pastebimos panašios tendencijos (žr. 3 priedą).

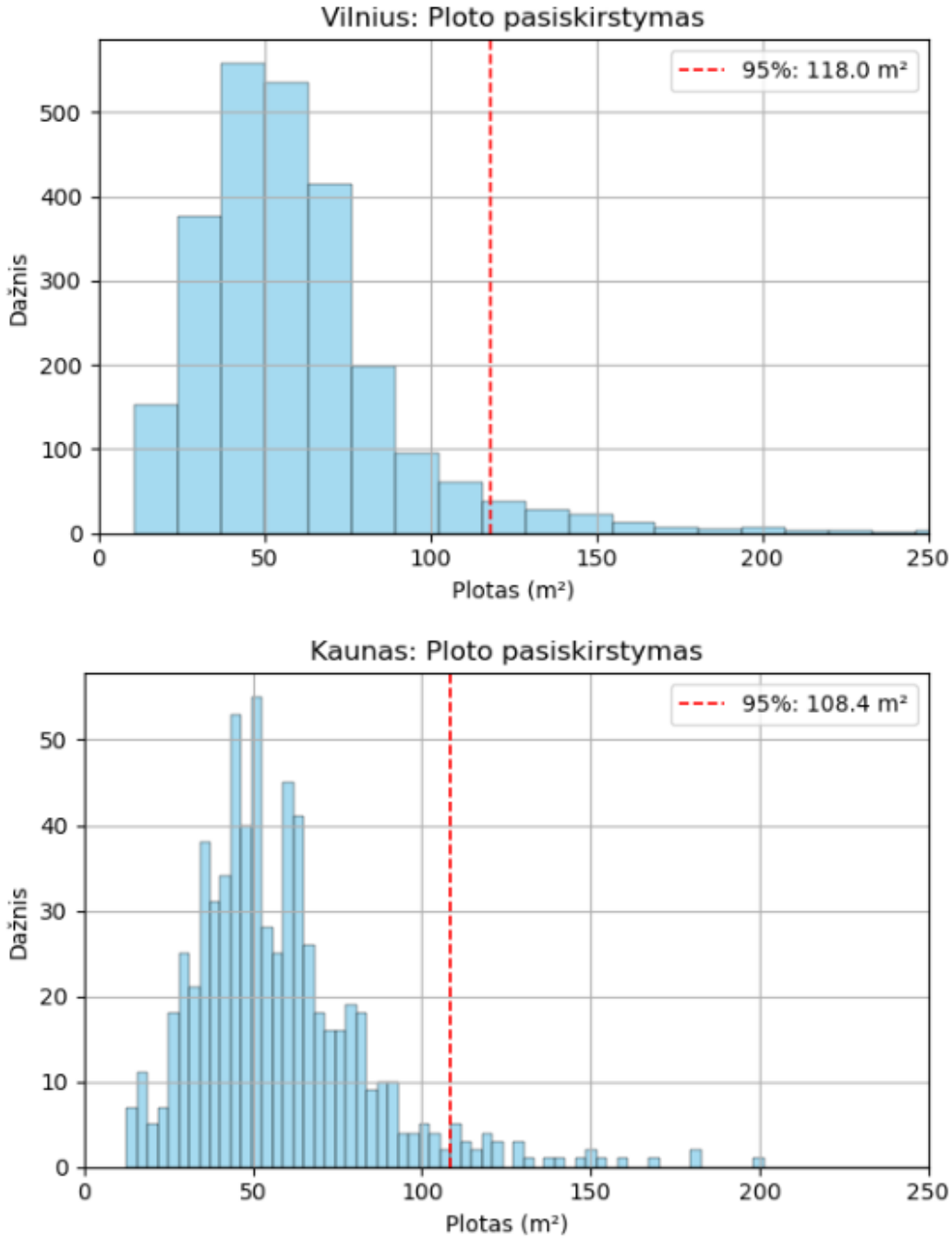
Perėjus prie išskirčių šalinimo, verta paminėti, kad daugelyje nagrinėtų tyrimų dažniausiai naudota standartinė anomalijų aptikimo priemonė – interkvartilinis atstumas (angl. *interquartile range*). Šis metodas leidžia identifikuoti ir pašalinti stebėjimus, kurie yra už nustatytų kvartilinių ribų. Šiame tyrime *IQR* metodas nebuvo naudotas, nes modeliavimo etape naudinga turėti ir labai pigius arba itin mažo ploto butus, kurie būtų pašalinti taikant *IQR* metodą. Dėl šios priežasties pasirinkta alternatyvi strategija – šalinti išskirtis remiantis imčių procentiliais. Tokiu būdu užtikrinta, kad būtų pašalintos tik pačios kraštutiniausios reikšmės (prabangūs butai, kurių kainos ar plotai yra neproporcingai dideli), taip pat išlaikyti tyrimui svarbūs mažiausios kainos arba mažiausio ploto butai. Pagal pateiktą butų pardavimo kainos skirstinių vizualizaciją galima pastebėti, kad tiek Vilniaus, tiek Kauno atveju

skirstiniai yra asimetriški ir turi ilgas dešiniąsias uodegas (žr. 13 pav). Atlikus analizę, nustatyta, kad Vilniaus butų pardavimo kainos 95-ojo procentilio reikšmė yra 540000 €, o Kauno – 345400 €.




**13 pav.** Vilniaus ir Kauno butų pardavimo kainos skirstiniai su atitinkamomis 95 procentilių reikšmėmis

Analizuojant ploto skirstinius, galima pastebėti, kad jie taip pat turi ilgas dešiniąsias uodegas (žr. 14 pav). Analogiškai kaip ir su pardavimo kainos skirstiniais, buvo nustatyta 95-ojo procentilio riba, pagal kurią buvo šalinami stebėjimai. Vilniaus atveju ši riba buvo 118 $m^2$, o Kauno – 108,4 $m^2$.




**14 pav.** Vilniaus ir Kauno butų ploto skirstiniai su atitinkamomis 95 procentilių reikšmėmis

Analizuojant duomenų rinkinius, buvo nagrinėti ir kambarių skaičiaus skirstiniai. Vilniaus butų pardavimo duomenyse visi stebėjimai pateko į 1-5 kambarių skaičiaus intervalą, todėl reikšmingos išskirtys nebuvo identifikuotos, o Kauno atveju duomenyse aptiktas įrašas su 8 kambariais, kuris buvo pašalintas iš imties. Įgyvendinus visas procedūras, Vilniaus butų pardavimo duomenų rinkinyje iš viso buvo pašalinti 174 stebėjimai, o Kauno – 55. Toliau likę kategoriniai kintamieji buvo išplėsti, naudojant anksčiau minėtą „one-hot“ kodavimą, taip pat, siekiant turėti kuo tvarkingesnius duomenis, suvienodintas visų kategorinių kintamųjų tipas. Žvalgomosios analizės metu modifikuoti butų pardavimo duomenų rinkiniai ir jų svarbiausių požymių empirinės charakteristikos pateikiamos 4 ir 5 lentelėse.

**4 lentelė.** Sutvarkyto Vilniaus butų pardavimo duomenų rinkinio svarbiausių požymių empirinės charakteristikos

| | Stebėjimai | Vidurkis | SD | Min | 25% | Mediana | 75% | Maks |
|---|---|---|---|---|---|---|---|---|
| **Kaina (€)** | 2353.0 | 186611.66 | 94457.07 | 21000.00 | 123000.00 | 166000.0 | 233000.0 | 539400.00 |
| **Plotas ($m^2$)** | 2353.0 | 53.93 | 20.58 | 10.86 | 39.92 | 52.0 | 67.0 | 118.00 |
| **Plotas vienam kambariui ($m^2$)** | 2353.0 | 24.46 | 6.08 | 6.00 | 20.48 | 23.5 | 27.5 | 78.18 |
| **Kambarių skaičius** | 2353.0 | 2.26 | 0.85 | 1.00 | 2.00 | 2.0 | 3.0 | 5.00 |
| **Pastato amžius (m.)** | 2353.0 | 33.83 | 28.44 | -1.00 | 8.00 | 25.0 | 55.0 | 275.00 |
| **Aukštas** | 2353.0 | 3.50 | 2.63 | 0.00 | 2.00 | 3.0 | 5.0 | 22.00 |
| **Aukštų skaičius** | 2353.0 | 6.05 | 3.59 | 1.00 | 4.00 | 5.0 | 7.0 | 29.00 |
| **Ats. iki centrinio taško kainos atžvilgiu (km)** | 2353.0 | 4.06 | 2.59 | 0.06 | 1.86 | 3.8 | 6.0 | 15.36 |
| **Ats. iki Vilniaus arkikatedros automobiliu (km)** | 2353.0 | 6.14 | 3.54 | 0.20 | 3.40 | 5.4 | 8.6 | 19.50 |
| **Ats. iki g. stoties automobiliu (km)** | 2353.0 | 7.10 | 4.17 | 0.10 | 3.20 | 7.0 | 10.5 | 22.40 |

**5 lentelė.** Sutvarkyto Kauno butų pardavimo duomenų rinkinio svarbiausių požymių empirinės charakteristikos

| | Stebėjimai | Vidurkis | SD | Min | 25% | Mediana | 75% | Maks |
|---|---|---|---|---|---|---|---|---|
| **Kaina (€)** | 625.0 | 129722.06 | 64329.67 | 19990.00 | 84999.00 | 119500.00 | 160000.00 | 345000.00 |
| **Plotas ($m^2$)** | 625.0 | 52.76 | 18.31 | 12.53 | 39.68 | 50.45 | 64.35 | 103.00 |
| **Plotas vienam kambariui ($m^2$)** | 625.0 | 23.42 | 5.54 | 12.25 | 20.00 | 22.30 | 25.68 | 50.44 |
| **Kambarių skaičius** | 625.0 | 2.32 | 0.81 | 1.00 | 2.00 | 2.00 | 3.00 | 5.00 |
| **Pastato amžius (m.)** | 625.0 | 48.66 | 33.84 | 0.00 | 30.00 | 48.00 | 60.00 | 236.00 |
| **Aukštas** | 625.0 | 3.40 | 2.33 | 0.00 | 2.00 | 3.00 | 4.00 | 16.00 |
| **Aukštų skaičius** | 625.0 | 5.15 | 2.76 | 1.00 | 3.00 | 5.00 | 5.00 | 16.00 |
| **Ats. iki centrinio taško kainos atžvilgiu (km)** | 625.0 | 3.29 | 1.82 | 0.18 | 1.83 | 3.44 | 4.45 | 9.64 |
| **Ats. iki Kauno Pilies automobiliu (km)** | 625.0 | 6.08 | 2.97 | 0.20 | 3.80 | 6.30 | 7.80 | 16.50 |
| **Ats. iki g. stoties automobiliu (km)** | 625.0 | 5.95 | 3.06 | 0.10 | 3.40 | 5.60 | 8.10 | 18.10 |

## 3.2. Optimalių struktūrinių parametrų paieška

Struktūrinių parametrų paieškai naudoti sutvarkyti duomenų rinkiniai ir anksčiau minėti „atstovaujantys“ modeliai – klasikiniams modeliams pasirinkti *Lasso*, *Ridge* ir *ElasticNet*, o medžių modeliams – atsitiktinis miškas. Reikėtų atkreipti dėmesį į tai, kad kokybės rodiklių skaičiavimas atliktas sugrąžinant prognozuojamą kintamąjį į standartinę reikšmę – kvadratinio metro kaina buvo dauginama iš ploto, o logaritmuotos prognozės konvertuotos naudojant eksponentinę funkciją. Vertinant tiesinius modelius, galima matyti, kad geriausias rezultatas, naudojant Vilniaus butų pardavimo duomenis, gautas su *Ridge* modeliu (pagal *MAPE* rodiklį). Akivaizdu, kad pagrindinių komponenčių analizė nėra naudinga, tuo tarpu kategoriniai požymiai yra visada naudingi. Taip pat pastebėta, kad efektyvesnis priklausomojo kintamojo pasirinkimas butų pardavimo duomenų atžvilgiu buvo kvadratinio metro kaina (žr. 6, 8 lenteles).

**6 lentelė.** Tiesinės regresijos modelių 10 geriausių struktūrinių parametrų sprendimų rezultatai, gauti naudojant Vilniaus butų pardavimo duomenų rinkinį

| Modelis | Priklausomas kint. | Ar naudojamas PCA? | Skalės transformacija | Ar naudojami kategoriniai požymiai ? | MAPE reikšmė (%) |
|---|---|---|---|---|---|
| Ridge | Log(kv_kaina) | Ne | Standard | Taip | **14,5901** |
| Ridge | Log(kv_kaina) | Ne | Robust | Taip | 14,5921 |
| Ridge | Log(kv_kaina) | Ne | Nenaudojama | Taip | 14,6416 |
| Ridge | Log(kv_kaina) | Ne | MinMax | Taip | 14,6698 |
| Ridge | kv_kaina | Ne | Robust | Taip | 15,2800 |
| Ridge | kv_kaina | Ne | Stardard | Taip | 15,2812 |
| Ridge | kv_kaina | Ne | MinMax | Taip | 15,2859 |
| Ridge | kv_kaina | Ne | Nenaudojama | Taip | 15,3257 |
| Lasso | kv_kaina | Ne | Robust | Taip | 15,3843 |
| Lasso | kv_kaina | Ne | Standard | Taip | 15,3870 |

Toliau nagrinėta papildomų požymių transformacijų įtaka modelių tikslumui. Tikslas – išsiaiškinti, ar skaitinių požymių sąveikos bei tolydžiųjų požymių skirstinių transformacijos padeda pagerinti modeliavimo rezultatus. Šiame etape pasirinktas geriausias modelis pagal 6 lentelės rezultatus (*Ridge*) su geriausiais struktūriniais parametrais.

**7 lentelė.** *PolynomialFeatures* ir *PowerTransform* funkcijų taikymo rezultatai, gauti naudojant *Ridge* modelį ir Vilniaus butų pardavimo duomenų rinkinį

| Ar naudojama ***PolynomialFeatures*** **funkcija?** | Ar naudojama ***PowerTransform*** **funkcija?** | MAPE reikšmė (%) |
|---|---|---|
| Taip | Taip | **13,5188** |
| Taip | Ne | 13,5571 |
| Ne | Taip | 14,4440 |
| Ne | Ne | 14,6932 |

Remiantis 7 lentelės rezultatais, galima pastebėti, kad geriausią prognozavimo tikslumą (mažiausią *MAPE* reikšmę) pavyko pasiekti tuomet, kai buvo taikomos tiek *PolynomialFeatures*, tiek *PowerTransform* funkcijos. Taip pat verta pažymėti, kad netaikant nei vienos funkcijos, *MAPE* reikšmė pakilo iki 14,69%, o tai patvirtina, jog šios funkcijos yra naudingos konstruojant regresijos modelius (naudojant Vilniaus butų pardavimo duomenis).

**8 lentelė.** Atsitiktinio miško modelio 10 geriausių struktūrinių parametrų sprendimų rezultatai, gauti naudojant Vilniaus butų pardavimo duomenų rinkinį

| Modelis | Priklausomas kint. | Ar naudojamas PCA? | Skalės transformacija | Ar naudojami kategoriniai požymiai ? | MAPE reikšmė (%) |
|---|---|---|---|---|---|
| RandomForest | kv_kaina | Ne | Robust | Taip | **12,3176** |
| RandomForest | kv_kaina | Ne | MinMax | Taip | 12,3210 |
| RandomForest | kv_kaina | Ne | Nenaudojama | Taip | 12,3225 |
| RandomForest | kv_kaina | Ne | Standard | Taip | 12,3225 |
| RandomForest | kv_kaina | Ne | MixMax | Ne | 12,4176 |
| RandomForest | kv_kaina | Ne | Nenaudojama | Ne | 12,4195 |

| | | | | | |
|---|---|---|---|---|---|
| RandomForest | kv_kaina | Ne | Robust | Ne | 12,4237 |
| RandomForest | kv_kaina | Ne | Standard | Ne | 12,4257 |
| RandomForest | kaina | Ne | Standard | Taip | 13,7170 |
| RandomForest | kaina | Ne | Robust | Taip | 13,7187 |

Atlikus keletą testavimų, išsiaiškinta, kad sprendimų medžių ansamblių modeliai nėra jautrūs tolydžiųjų požymių skirstiniams. Dėl šios priežasties sprendimų medžių modeliams pasirinkta nagrinėti tik *PolynomialFeatures* funkcijos naudą. Gauti rezultatai parodė, kad polinominės požymių sąveikos padėjo pagerinti prognozavimo tikslumą ir šių modelių atžvilgiu.

Remiantis tokia pačia procedūra, atlikta ir Kauno butų pardavimo duomenų geriausių struktūrinių parametrų paieška, toliau pateikti paieškos rezultatai (žr. 9, 11 lenteles).

**9 lentelė.** Tiesinės regresijos modelių 10 geriausių struktūrinių parametrų sprendimų rezultatai, gauti naudojant Kauno butų pardavimo duomenų rinkinį

| **Modelis** | **Priklausomas kint.** | **Ar naudojamas PCA?** | **Skalės transformacija** | **Ar naudojami kategoriniai požymiai ?** | **MAPE reikšmė (%)** |
|---|---|---|---|---|---|
| Ridge | Log(kv_kaina) | Ne | Nenaudojama | Taip | **18,0763** |
| Ridge | Log(kv_kaina) | Ne | MinMax | Taip | 18,1280 |
| Ridge | Log(kv_kaina) | Ne | Robust | Taip | 18,1375 |
| Ridge | Log(kv_kaina) | Ne | Standard | Taip | 18,1375 |
| Lasso | kv_kaina | Ne | Nenaudojama | Taip | 18,6579 |
| Ridge | kv_kaina | Ne | Nenaudojama | Taip | 18,7008 |
| Lasso | kv_kaina | Ne | MinMax | Taip | 18,7013 |
| Ridge | kv_kaina | Ne | MinMax | Taip | 18,7165 |
| Lasso | kv_kaina | Ne | Standard | Taip | 18,7259 |
| Lasso | kv_kaina | Ne | Robust | Taip | 18,7334 |

Skirtingai nei Vilniaus atveju, Kauno butų pardavimo rinkiniui taikytos *PolynomialFeatures* ir *PowerTransform* funkcijos nepadėjo pagerinti *Ridge* modelio prognozavimo tikslumo – mažiausia *MAPE* reikšmė (18,54%) gauta tada, kai nebuvo taikoma nė viena iš minėtų funkcijų (žr. 10 lentelę). Toks rezultatas gali būti paaiškintas sąlyginai mažu Kauno butų pardavimo duomenų rinkinio stebėjimų skaičiumi, akivaizdu, kad požymių transformacijų funkcijos „užkimšo“ *Ridge* modelį neinformatyviais požymiais bei pablogino modelio apibendrinimo gebėjimą.

**10 lentelė.** *PolynomialFeatures* ir *PowerTransform* funkcijų taikymo rezultatai, gauti naudojant *Ridge* modelį ir Kauno butų pardavimo duomenų rinkinį

| **Ar naudojama *PolynomialFeatures* funkcija ?** | **Ar naudojama *PowerTransform* funkcija ?** | **MAPE reikšmė (%)** |
|---|---|---|
| Ne | Ne | **18,5403** |
| Ne | Taip | 18,5793 |
| Taip | Ne | 21,8825 |
| Taip | Taip | 21,9998 |

**11 lentelė.** Atsitiktinio miško modelio 10 geriausių struktūrinių parametrų sprendimų rezultatai, gauti naudojant Kauno butų pardavimo duomenų rinkinį

| Modelis | Priklausomas kint. | Ar naudojamas PCA? | Skalės transformacija | Ar naudojami kategoriniai požymiai ? | MAPE reikšmė (%) |
|---|---|---|---|---|---|
| RandomForest | kv_kaina | Ne | Nenaudojama | Taip | **17,5178** |
| RandomForest | kv_kaina | Ne | Robust | Taip | 17,5280 |
| RandomForest | kv_kaina | Ne | MinMax | Taip | 17,5288 |
| RandomForest | kv_kaina | Ne | Standard | Taip | 17,5440 |
| RandomForest | kv_kaina | Ne | MixMax | Ne | 18,4381 |
| RandomForest | kv_kaina | Ne | Robust | Ne | 18,4406 |
| RandomForest | kv_kaina | Ne | Standard | Ne | 18,4453 |
| RandomForest | kv_kaina | Ne | Nenaudojama | Ne | 18,4460 |
| RandomForest | kaina | Ne | MinMax | Taip | 19,6963 |
| RandomForest | kaina | Ne | Standard | Taip | 19,7070 |

Dar kartą patikrinus polinominių požymių sąveikų naudą medžių modeliams, išsiaiškinta, kad skirtingai nei Vilniaus atveju, naudojant Kauno butų pardavimo duomenis šios sąveikos nepadėjo pagerinti prognozavimo tikslumo.

### 3.3. Statistiškai reikšmingų požymių atranka

Toliau pateikiami modeliavimo etape naudoti galutiniai paaiškinamųjų požymių rinkiniai, gauti naudojant skirtingas atrankos strategijas. Svarbu paminėti, kad nepaisant anksčiau gautų papildomų požymių transformacijų naudingumo rezultatų, priimtas sprendimas ir tiesinės regresijos, ir sprendimų medžių ansamblių modeliams naudoti *PolynomialFeatures* funkcijos pagalba papildytus duomenų rinkinius. Reguliarizuotų regresijų atrankos atveju išbandyti *Lasso* ir *Ridge* modeliai su automatiškai parinktu $\alpha$ hiperparametru. Siekiant nepašalinti per daug požymių, papildomai buvo eksperimentuota su reguliarizuotos regresijos modelių ribinių reikšmių įverčiais. *Boruta* algoritmu paremtai atrankai panaudotas atsitiktinio miško modelis su 100 medžių. Kaip ir buvo minėta metodologijos apžvalgoje, kuriant regresijos modelius reikėtų išmėginti keletą skirtingų požymių rinkinių, todėl papildomai sukurti ir mišrieji rinkiniai. 12, 13, 14, 15 lentelėse pateikiami originalūs (be *PolynomialFeatures* požymių) ir atrankos metu gauti butų pardavimo ir nuomos duomenų paaiškinamųjų požymių rinkiniai.

**12 lentelė.** Galutiniai Vilniaus butų pardavimo duomenų paaiškinamųjų požymių rinkiniai

| Išsaugoto failo pavadinimas | Požymių skaičius |
|---|---|
| Vilnius_org.csv | 80 |
| Vilnius_reg_atranka.csv | 58 |
| Vilnius_boruta.csv | 9 |
| Vilnius_misrus.csv | 64 |

**13 lentelė.** Galutiniai Kauno butų pardavimo duomenų paaiškinamųjų požymių rinkiniai

| Išsaugoto failo pavadinimas | Požymių skaičius |
|---|---|
| Kaunas_org.csv | 62 |
| Kaunas_reg_atranka.csv | 46 |
| Kaunas_boruta.csv | 8 |

| | |
|---|---|
| Kaunas_misrus.csv | 47 |

**14 lentelė.** Galutiniai Vilniaus butų nuomos duomenų paaiškinamųjų požymių rinkiniai

| **Išsaugoto failo pavadinimas** | **Požymių skaičius** |
|---|---|
| Vilnius_nuoma_org.csv | 75 |
| Vilnius_nuoma_reg_atranka.csv | 37 |
| Vilnius_nuoma_boruta.csv | 6 |
| Vilnius_nuoma_misrus.csv | 37 |

**15 lentelė.** Galutiniai Kauno butų nuomos duomenų paaiškinamųjų požymių rinkiniai

| **Išsaugoto failo pavadinimas** | **Požymių skaičius** |
|---|---|
| Kaunas_nuoma_org.csv | 58 |
| Kaunas_nuoma_reg_atranka.csv | 12 |
| Kaunas_nuoma_boruta.csv | 5 |
| Kaunas_nuoma_misrus.csv | 14 |

Atlikus atranką buvo pastebėta, kad egzistuoja tokie požymiai, kurie yra statistiškai reikšmingi prognozuojant tiek butų pardavimo, tiek nuomos kainą. Keletas pavyzdžių – bendras plotas, kambarių skaičius, papildomų patalpų skaičius, ypatybių skaičius bei kai kurie atstumų požymiai. Tačiau išryškėjo ir esminiai skirtumai. Į butų pardavimo kainos paaiškinamųjų požymių rinkinius pateko nemažai kategorinių požymių, nurodančių objekto rajoną, pastato tipą bei šildymo sistemą, taip pat keletas požymių, gautų naudojant *PolynomialFeatures* funkciją, o naudojant butų nuomos duomenis, didžioji dalis šių požymių pasirodė kaip statistiškai nereikšmingi. Galutinius butų nuomos kainos paaiškinamųjų požymių rinkinius iš esmės sudarė pagrindiniai skaitiniai kintamieji. Pavyzdžiui, nagrinėjant Kauno butų nuomos duomenų mišrųjį požymių rinkinį, pastebėta, kad iš kategorinių požymių statistiškai reikšmingi buvo tik 3 binariniai kintamieji – „rajonas_Centras“, „sildymas_centrinis“ ir „sildymas_kolektorinis“. Galima teigti, kad kuriant regresijos modelius pagal Vilniaus ir Kauno butų nuomos kainas, dėmesys turėtų būti skiriamas kiekybiniams kintamiesiems, o lokaciniai ar kokybiniai požymiai yra mažiau svarbūs.

### 3.4. Modelių hiperparametrų optimizavimas

Šiame etape atliktas išsamus modelių hiperparametrų optimizavimas su *Optuna* biblioteka, papildomai kiekvienam regresijos modeliui parinktas tinkamiausias paaiškinamųjų požymių rinkinys. Siekiant pagreitinti optimizavimo procesą, pritaikytas skaičiavimų lygiagretinimas tiems modeliams, kurie palaiko šią funkciją. Šioje tyrimo dalyje naudota įranga – 4,0 GHz Intel Core i7-6700K procesorius su 8 branduoliais ir 16 gigabaitų operatyviosios atminties. Atliekant pradinį testavimą kiekvienam rinkiniui skirta 15 hiperparametrų kombinacijų bandymų, tada rezultatai buvo vertinti pagal *MAPE*, tokiu būdu nusprendžiant, kuris paaiškinamųjų požymių rinkinys yra geriausias kiekvienam tyrime naudojamam modeliui. Atlikus pradinį testavimą, toliau atliktas kiekvieno regresijos modelio pagrindinis hiperparametrų optimizavimas, kuriam skirta 100 bandymų arba 30 minučių skaičiavimo laiko. Atliekant optimizavimą, kiekvienas modelis buvo dar kartą patikrintas pagal tai, ar verta logaritmuoti priklausomąjį kintamąjį. Šioje tyrimo dalyje visuose modeliuose naudotas vienodas priklausomas kintamasis – buto kvadratinio metro kaina, o gavus prognozes, modelių kokybės kriterijų skaičiavimas atliktas pagal anksčiau minėtą procedūrą, kai prieš vertinimą prognozės yra konvertuojamos į buto kainą. Visa tai užtikrino sklandžią hiperparametrų patikrinimo

procedūrą. Toliau pateikiami modelių hiperparametrų optimizavimo rezultatai, naudojant Vilniaus butų pardavimo duomenis (žr. 16 lentelę).

**16 lentelė.** Visų modelių hiperparametrų optimizavimo rezultatai, gauti naudojant Vilniaus butų pardavimo duomenų rinkinį

| **Modelis** | **Geriausias požymių rinkinys** | **Ar logaritmuoti „kv_kaina“ ?** | **Patikrintų hiperparametrų kombinacijų skaičius** | **Bendras skaičiavimo laikas** | **Geriausia MAPE reikšmė (%)** |
|---|---|---|---|---|---|
| LinearRegression | Mišrus | Taip | - | - | 14,11 |
| Ridge | Mišrus | Taip | 100 | 8,5 s | 14,13 |
| Lasso | Reg. atrankos | Taip | 100 | 17 s | 14,19 |
| ElasticNet | Reg. atrankos | Taip | 100 | 26 s | 14,20 |
| BayesianRidge | Mišrus | Taip | 100 | 18 s | 14,13 |
| SGD | Reg. atrankos | Taip | 100 | 24,4 s | 15,07 |
| SVR | Mišrus | Taip | 100 | 2 min 18 s | 12,19 |
| RandomForest | Mišrus | Ne | 100 | 7 min 51 s | 12,04 |
| GradientBoostingRegressor | Reg. atrankos | Ne | 55 | 30 min 23 s | 11,36 |
| BaggingRegressor | Originalus | Ne | 100 | 11 min 17 s | 11,96 |
| ExtraTreesRegressor | Boruta atrankos | Ne | 100 | 2 min 20 s | 12,50 |
| CatBoostRegressor | Originalus | Ne | 85 | 30 min 10 s | **11,10** |
| XGBoostRegressor | Originalus | Ne | 100 | 3 min 45 s | 11,18 |
| LGBMRegressor | Reg. atrankos | Ne | 100 | 3 min 3 s | 11,48 |
| HistGradientBoostingRegressor | Reg. atrankos | Ne | 100 | 19 min 53 s | 11,43 |
| AdaBoostRegressor | Mišrus | Taip | 66 | 30 min 17 s | 15,23 |
| KNeighborsRegressor | Mišrus | Taip | 100 | 20 s | 13,17 |

Galima pastebėti, kad visi klasikiniai regresijos modeliai pasižymėjo ypatingai trumpu skaičiavimo laiku. Akivaizdu, kad nors lygiagretūs skaičiavimai su visais branduoliais padeda pagreitinti optimizavimo procesą, bendras skaičiavimo laikas eksponentiškai didėja pagal regresijos modelyje esančių hiperparametrų skaičių. Kai kuriems modeliams (pvz., *CatBoostRegressor* arba *GradientBoostingRegressor*) nepavyko atlikti visų 100 bandymų per nustatytą 30 minučių skaičiavimo laiką, tačiau galima pastebėti, kad šie modeliai pasižymėjo pakankamai gerais *MAPE* rodiklio rezultatais net ir po 50 paieškos algoritmo iteracijų. Geriausias rezultatas užfiksuotas taikant *CatBoostRegressor* modelį (11,10% *MAPE* reikšmė). Visų modelių optimizavimas, naudojant

Vilniaus butų pardavimo duomenis, užtruko 2 valandas ir 32 minutes. Tokia pati optimizavimo procedūra atlikta naudojant Kauno butų pardavimo duomenis (žr. 17 lentelę).

**17 lentelė.** Visų modelių hiperparametrų optimizavimo rezultatai, gauti naudojant Kauno butų pardavimo duomenų rinkinį

| **Modelis** | **Geriausias požymių rinkinys** | **Ar logaritmuoti „kv_kaina“ ?** | **Patikrintų hiperparametrų kombinacijų skaičius** | **Bendras skaičiavimo laikas** | **Geriausia MAPE reikšmė (%)** |
|---|---|---|---|---|---|
| LinearRegression | Reg. atrankos | Taip | - | - | 17,81 |
| Ridge | Reg. atrankos | Taip | 100 | 4,2 s | 17,79 |
| Lasso | Originalus | Taip | 100 | 2,5 s | 17,67 |
| ElasticNet | Reg. atrankos | Ne | 100 | 5,7 s | 18,43 |
| BayesianRidge | Reg. atrankos | Taip | 100 | 7 s | 17,79 |
| SGD | Originalus | Taip | 100 | 6,6 s | 21,89 |
| SVR | Boruta atrankos | Taip | 100 | 15,8 s | 16,47 |
| RandomForest | Reg. atrankos | Ne | 100 | 2 min 19 s | 17,29 |
| GradientBoostingRegressor | Originalus | Ne | 100 | 5 min 9 s | **15,55** |
| BaggingRegressor | Reg. atrankos | Ne | 100 | 3 min 5 s | 17,32 |
| ExtraTreesRegressor | Mišrus | Ne | 100 | 1 min 53 s | 17,56 |
| CatBoostRegressor | Originalus | Ne | 100 | 28 min | 15,61 |
| XGBoostRegressor | Originalus | Ne | 100 | 1 min 56 s | 15,64 |
| LGBMRegressor | Originalus | Ne | 100 | 1 min 3 s | 16,64 |
| HistGradientBoostingRegressor | Originalus | Ne | 100 | 5 min 19 s | 16,41 |
| AdaBoostRegressor | Originalus | Taip | 100 | 7 min 13 s | 19,31 |
| KNeighborsRegressor | Originalus | Taip | 100 | 9 s | 18,86 |

Šį kartą visiems modeliams pavyko gauti 100 hiperparametrų kombinacijų bandymų rezultatus, modelių optimizavimas užtruko 58 minutes, o geriausias rezultatas, naudojant Kauno butų pardavimo duomenis, užfiksuotas taikant *GradientBoostingRegressor* modelį (15,55% *MAPE* reikšmė).

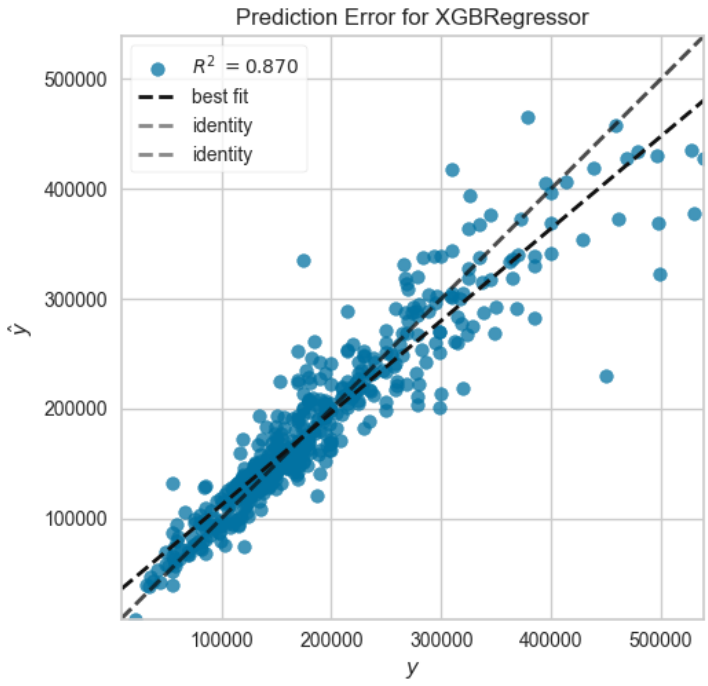
## 3.5. Geriausi butų pardavimo kainos prognozavimo modeliai

Toliau pateikiami 5 didžiausiu tikslumu pasižymėję modeliai (pagal *MAPE*), naudojant Vilniaus butų pardavimo duomenis. Pastebėta, kad optimizavimas padėjo pagerinti visų 5 modelių rezultatus, lyginant su baziniais modeliais, kurie turėjo standartinius hiperparametrus. *LGBMRegressor* modeliui užfiksuota net 10% geresnė *MAPE* rodiklio reikšmė (žr. 18 lentelę).

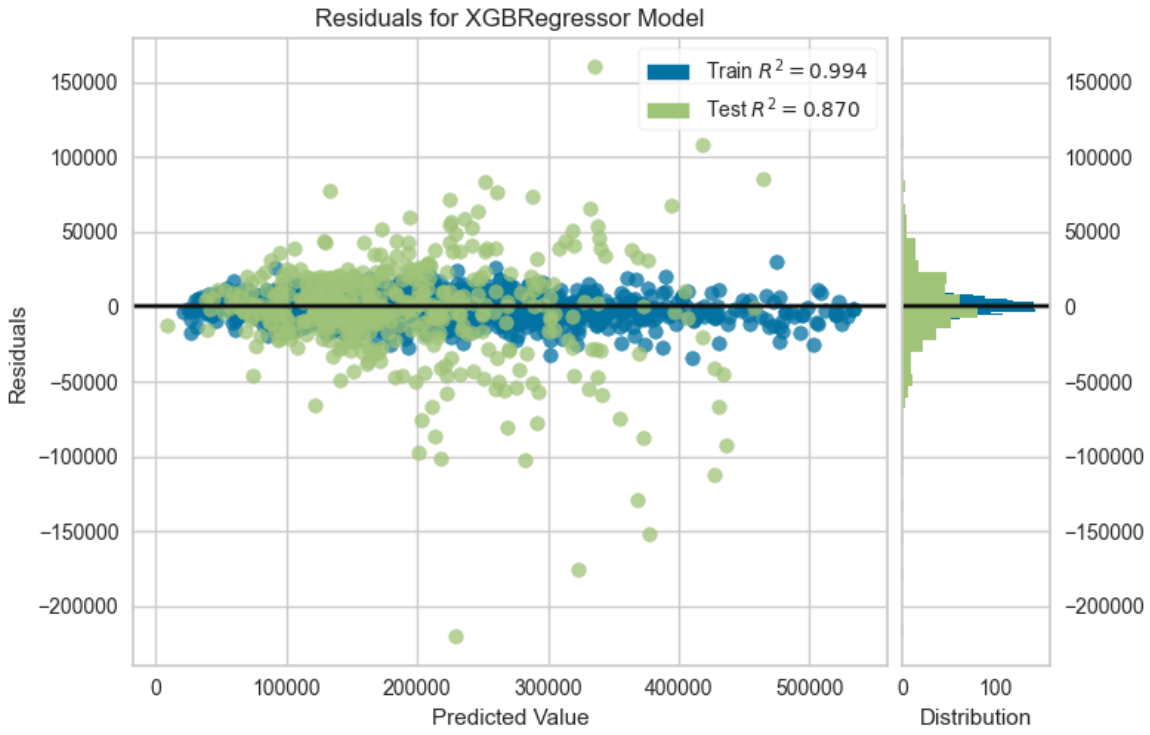
**18 lentelė.** 5 geriausi modeliai, gauti naudojant Vilniaus butų pardavimo duomenų rinkinį

| **Modelis** | **R^2** | **RMSE** | **MAPE (%)** | **MAPE pokytis lyginant su baziniu modeliu (%)** |
|---|---|---|---|---|
| CatBoostRegressor | 0,8960 | 30351,74 | 11,10 | 1,99 |
| XGBoostRegressor | 0,8944 | 30539,57 | 11,30 | 5,35 |
| GradientBoostingRegressor | 0,8891 | 31370,33 | 11,48 | 0,68 |
| HistGradientBoostingRegressor | 0,8892 | 31338,14 | 11,42 | 2,45 |
| LGBMRegressor | 0,8936 | 30722,83 | 11,35 | 10,83 |

Siekiant praplėsti modeliavimo rezultatus, papildomai pavaizduotas geriausio modelio prognozuotų ir tikrųjų reikšmių grafikas. Kadangi *CatBoostRegressor* nėra suderintas su *yellowbrick* liekanų atvaizdavimo biblioteka, Vilniaus atveju vizualizacijoms naudotas antrasis pagal tikslumą modelis *MAPE* rodiklio atžvilgiu – *XGBoostRegressor*. Kartu su prognozuotų ir tikrųjų reikšmių grafiku pateikiamas šio modelio liekanų pasiskirstymo grafikas (žr. 15, 16 pav.).




**15 pav.** Prognozuotų ir tikrųjų reikšmių grafikas, gautas naudojant Vilniaus butų pardavimo duomenis ir *XGBoostRegressor* modelį



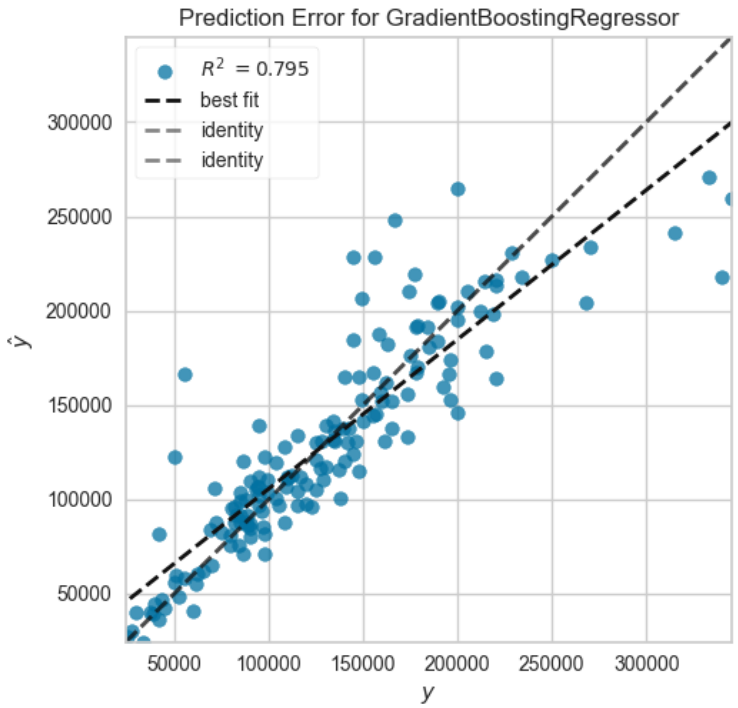

**16 pav.** Liekanų pasiskirstymo grafikas, gautas naudojant Vilniaus butų pardavimo duomenis ir *XGBoostRegressor* modelį

Remiantis 15 pav. pateiktu grafiku, galima pastebėti, kad *XGBoostRegressor* modelis pakankamai gerai prognozavo vidutinės ir mažos vertės butų pardavimo kainas – dauguma taškų 0-200 tūkstančių eurų intervale yra glaudžiai išsidėstę aplink tapatybės liniją. Vis dėlto, galima matyti, kad modelis prasčiau susitvarkė su brangesniais objektais – pradedant nuo 300 tūkstančių eurų, procentinės paklaidos tapo vis didesnės, tai patvirtina ir liekanų pasiskirstymo grafikas.

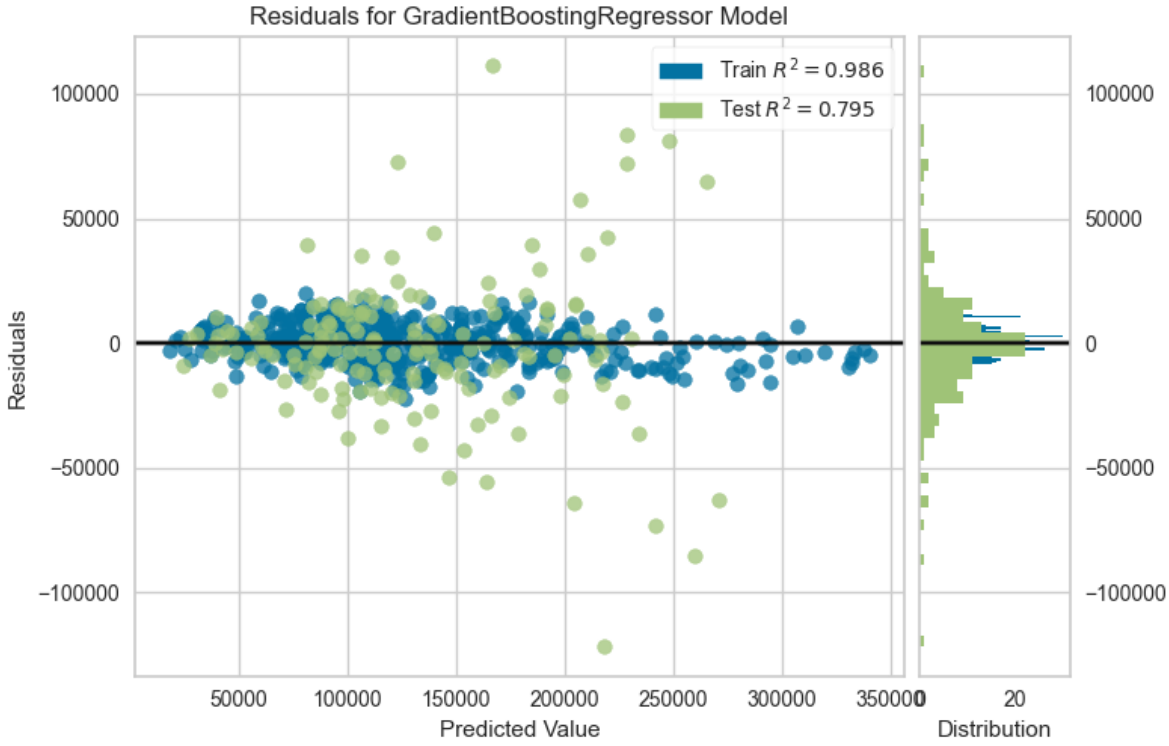
Toliau pateikiami geriausi modeliavimo rezultatai, naudojant Kauno butų pardavimo duomenų rinkinį, taip pat prognozuotų ir tikrųjų reikšmių bei liekanų pasiskirstymo grafikai, gauti naudojant didžiausiu tikslumu pasižymėjusį modelį – *GradientBoostingRegressor* (žr. 19 lentelę ir 17, 18 pav.).

**19 lentelė.** 5 geriausi modeliai, gauti naudojant Kauno butų pardavimo duomenų rinkinį

| **Modelis** | **R^2** | **RMSE** | **MAPE (%)** | **MAPE pokytis lyginant su baziniu modeliu (%)** |
|---|---|---|---|---|
| GradientBoostingRegressor | 0,8354 | 25835,27 | 15,54 | 4,26 |
| CatBoostRegressor | 0,8363 | 25809,17 | 15,60 | 1,53 |
| XGBoostRegressor | 0,8282 | 26439,45 | 15,63 | 9,46 |
| HistGradientBoostingRegressor | 0,8108 | 27833,07 | 16,41 | 4,38 |
| SVR | 0,8242 | 26745,54 | 16,47 | 17,05 |



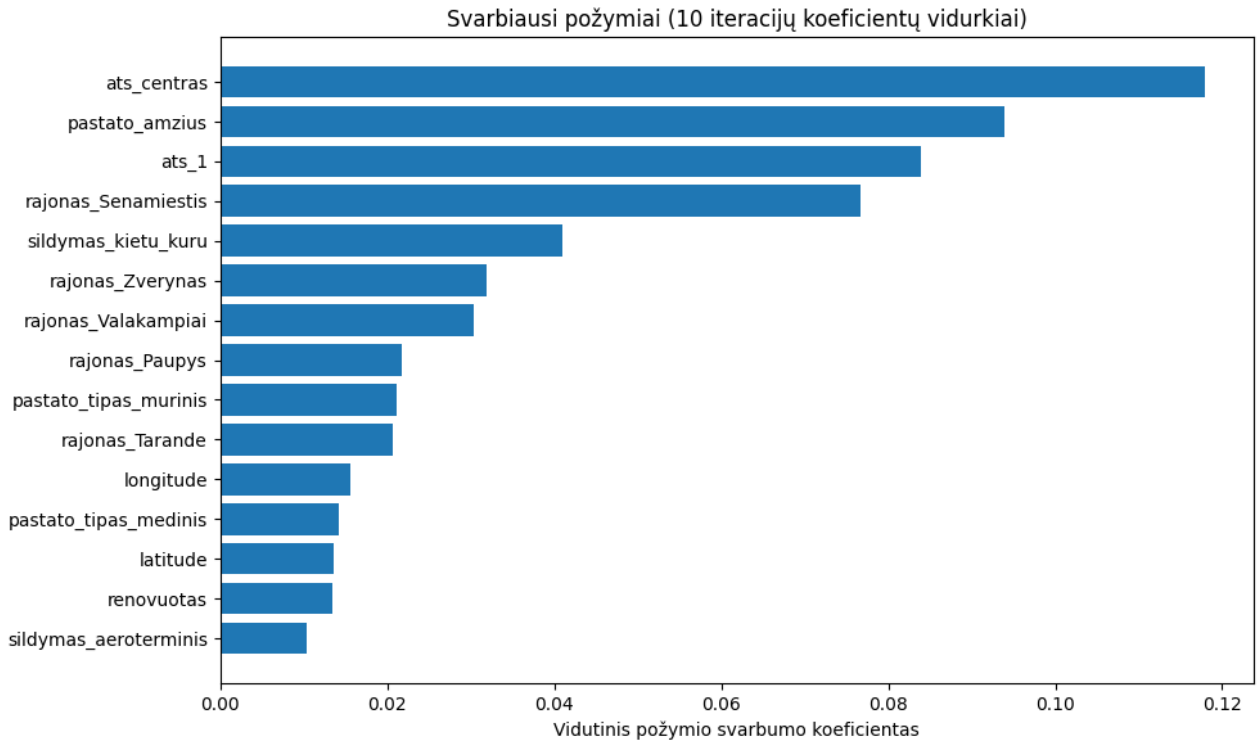

**17 pav.** Prognozuotų ir tikrųjų reikšmių grafikas, gautas naudojant Kauno butų pardavimo duomenis ir *GradientBoostingRegressor* modelį



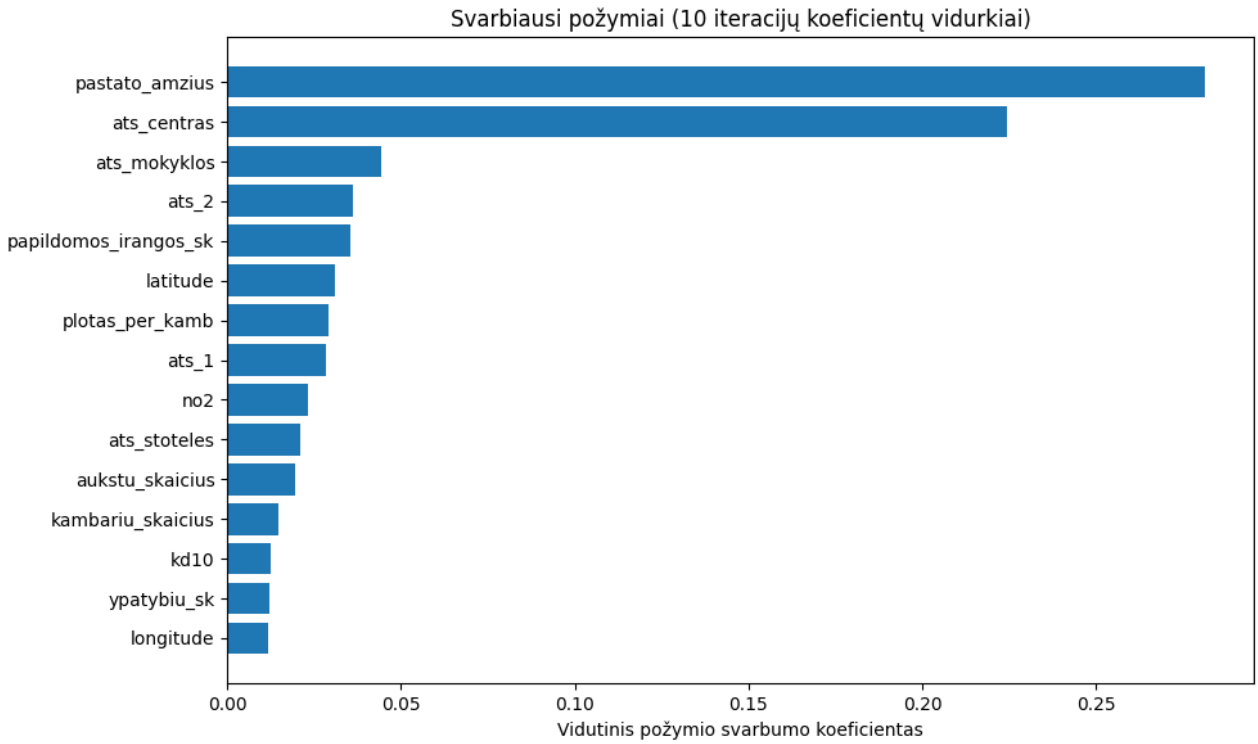

**18 pav.** Liekanų pasiskirstymo grafikas, gautas naudojant Kauno butų pardavimo duomenis ir *GradientBoostingRegressor* modelį

Nepaisant to, kad Kauno butų pardavimo duomenų rinkinį sudarė tik 625 stebėjimai, galima pastebėti, kad *GradientBoostingRegressor* modelis pakankamai gerai prognozavo vidutinės ir mažos vertės butų kainas. Vis dėlto, liekanų pasiskirstymo grafikas rodo sąlyginai didelį prognozavimo neapibrėžtumą, kuris ypatingai pastebimas kraštutinėse stebėjimų imties reikšmėse.

Taip pat nagrinėta, kurie paaiškinamieji požymiai modeliams yra svarbiausi. Šiame etape panaudoti geriausi modeliai su optimaliais hiperparametrais, o jų apmokymui skirta 10 iteracijų, kiekvieną kartą atliekant atsitiktinį treniravimo ir testavimo duomenų išskaidymą. Reikėtų atkreipti dėmesį į tai, kad naudojant butų pardavimo kainą kaip priklausomąjį kintamąjį, susiduriama su problema, kai plotas tampa svarbiausiu paaiškinamuoju požymiu, kas ir taip yra akivaizdu. Siekiant gauti informatyvesnius rezultatus, šioje tyrimo dalyje pasirinkta naudoti butų kvadratinio metro kainą, tokiu būdu eliminuota ploto požymio įtaka. Po kiekvienos iteracijos fiksuoti svarbumo koeficientai, kurie apskaičiuoti pagal kiekvieno požymio indėlį į sprendimų medžių struktūroje atliekamus skaidymus. Paprastai tariant, kuo dažniau konkretus požymis yra naudojamas sprendimų medžių šakose ir kuo labiau jis sumažina paklaidą kiekviename padalijime, tuo svarbesnis jis yra skaičiuojant galutinę prognozę. Atlikus 10 algoritmo iteracijų, apskaičiuotas kiekvieno požymio vidutinis svarbumo koeficientas. Vizualizacijų kūrimui atrinkta 15 svarbiausių paaiškinamųjų požymių, išrikiuotų pagal svarbumo koeficientus (žr. 19, 20 pav.).




**19 pav.** Paaiškinamieji požymiai ir jų svarbumo koeficientai, gauti naudojant Vilniaus butų pardavimo duomenis ir *XGBoostRegressor* modelį




**20 pav.** Paaiškinamieji požymiai ir jų svarbumo koeficientai, gauti naudojant Kauno butų pardavimo duomenis ir *GradientBoostingRegressor* modelį

Pastebėta, kad tiek Vilniaus, tiek Kauno atveju svarbiausi buto kvadratinio metro pardavimo kainos paaiškinamieji požymiai buvo tokie patys – pastato amžius bei Euklido atstumas iki miesto centro. Yra ir nemažai skirtumų – Vilniaus atveju svarbūs pasirodė kategoriniai požymiai, nurodantys rajoną, pastato tipą bei šildymo tipą, o Kauno atveju svarbesni buvo atstumų požymiai ir kiti skaitiniai kintamieji, pavyzdžiui, papildomos įrangos skaičius, buto koordinatės ir pan.

### 3.6. Geriausi butų nuomos kainos prognozavimo modeliai

Kuriant nuomos duomenimis paremtus modelius, nuosekliai vykdyti visi anksčiau išvardinti procesai:

- žvalgomosios analizės, požymių inžinerijos ir išskirčių šalinimo procedūros;
- optimalių struktūrinių parametrų paieška;
- statistiškai reikšmingų požymių atranka (žr. 14 ir 15 lentelių rezultatus);
- modelių hiperparametrų optimizavimas (žr. 4 ir 5 priedų rezultatus).

Butų nuomos rinkiniai buvo papildyti anksčiau sukurtais paaiškinamaisiais požymiais, taip pat, remiantis *KDE* metodu, perskaičiuoti „centriniai“ taškai ir Euklido atstumai. Optimalių struktūrinių parametrų paieškos rezultatai iš esmės buvo identiški, lyginant su tais, kurie buvo gauti naudojant butų pardavimo duomenų rinkinius. Geriausias priklausomas kintamasis visais atvejais buvo buto kvadratinio metro nuomos kaina, pagrindinių komponenčių analizė neteikė naudos, o kategoriniai požymiai visais atvejais buvo naudingi. Visų modelių optimizavimas, naudojant Vilniaus butų nuomos duomenis, užtruko 57 minutes (be *CatBoostRegressor*, kuris dėl neaiškios priežasties nesuveikė), o naudojant Kauno butų nuomos duomenis – 49 minutes. Toliau pateikiami 5 didžiausiu tikslumu pasižymėję modeliai (žr. 20 ir 21 lenteles). Atlikus detalesnę analizę, dar kartą įsitikinta, kad hiperparametrų optimizavimas padeda pagerinti prognozių tikslumą. Nors pagal *MAPE* rodiklio reikšmes gauti rezultatai neatrodė prastai, verta pažymėti, kad geriausių modelių $R^2$ rodikliai buvo

gerokai mažesni nei butų pardavimo duomenų atveju – 0,77 naudojant *GradientBoostingRegressor* ir Vilniaus butų nuomos duomenis ir tik 0,69 naudojant *CatBoostRegressor* ir Kauno butų nuomos duomenis. Galima susidaryti nuomonę, kad nuomos kainų modeliavimas yra sudėtingesnis uždavinys, kuriam spręsti reikėtų dar gilesnės analizės bei alternatyvių paaiškinamųjų požymių.

**20 lentelė.** 5 geriausi modeliai, gauti naudojant Vilniaus butų nuomos duomenų rinkinį

| **Modelis** | **R^2** | **RMSE** | **MAPE (%)** | **MAPE pokytis lyginant su baziniu modeliu (%)** |
|---|---|---|---|---|
| GradientBoostingRegressor | 0,7781 | 96,20 | 13,56 | 4,19 |
| XGBoostRegressor | 0,7782 | 96,13 | 13,72 | 3,70 |
| RandomForestRegressor | 0,7709 | 97,65 | 14,07 | 2,00 |
| BaggingRegressor | 0,7658 | 98,79 | 14,15 | 7,00 |
| SVR | 0,7252 | 106,94 | 14,37 | 4,99 |

**21 lentelė.** 5 geriausi modeliai, gauti naudojant Kauno butų nuomos duomenų rinkinį

| **Modelis** | **R^2** | **RMSE** | **MAPE (%)** | **MAPE pokytis lyginant su baziniu modeliu (%)** |
|---|---|---|---|---|
| CatBoostRegressor | 0,6966 | 86,65 | 16,32 | 0,61 |
| XGBoostRegressor | 0,6922 | 86,94 | 16,56 | 5,92 |
| GradientBoostingRegressor | 0,6745 | 89,69 | 16,74 | 3,49 |
| BaggingRegressor | 0,6856 | 88,21 | 17,08 | 1,21 |
| ExtraTreesRegressor | 0,6759 | 89,27 | 17,10 | -0,88 |

## 3.7. Pardavimo duomenimis paremtų modelių praktinis taikymas

Toliau pateikiamas regresijos modelių praktinio taikymo pavyzdys, kuriame buvo naudotas 5 geriausių regresijos modelių prognozių vidurkis. Šiame pavyzdyje duomenų rinkiniai buvo segmentuoti pagal svarbiausius požymius pirkėjo atžvilgiu – kambarių skaičiaus, pastato statybos metų (pastato amžiaus), atstumo iki miesto centro. Siekiant pateikti praktinio taikymo pavyzdį, kuris yra labiau aktualus mažmeniniam investuotojui arba pirmąjį butą perkančiam gyventojui, pardavimo duomenyse apibrėžti globalūs ribojimai – vertinti tik 1, 2, 3 kambarių butai, kurių pardavimo kaina neviršija 205 tūkst. eurų. Tokių, kriterijus atitinkančių alternatyvų Vilniaus butų pardavimo duomenų rinkinyje buvo 1485, o Kauno – 519. Modeliavimas atliktas treniruojant geriausius modelius su pilnais duomenų rinkiniais, iš kurių buvo pašalintas 1 testavimo segmentas (pvz., 1 kambario butai, kurių pastato amžiaus požymio reikšmė yra 0-5 intervale), vėliau įvertinti prognozių rezultatai, prognozės anuliuotos ir procedūra kartota naudojant sekančio segmento duomenis. Tokiu būdu bandyta suprasti, kaip regresijos modeliai vertina konkretų rinkos segmentą ir kokia procentinė dalis atitinkamo segmento stebėjimų yra pervertinta (kai modelių prognozuojama buto pardavimo kaina yra didesnė už faktinę pardavimo kainą). Reikia atsižvelgti ir į tai, kad kai kurių stebėjimų paklaidos gali būti didesnės nei pagal *MAPE* rodiklį, todėl šiam rinkos vertinimo procesui atrinkti tik tie stebėjimai, kurių prognozuojama pardavimo kaina skyrėsi ne daugiau kaip 20% nuo faktinės kainos.

Modeliavimo pagal segmentus rezultatai Vilniaus butų pardavimo duomenų rinkiniui pateikti 22, 23, 24 lentelėse.

**22 lentelė.** Vilniaus butų pardavimo kainų modeliavimo rezultatai, gauti naudojant kambarių skaičiaus ir pastato amžiaus požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 0-5 metų intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 6-25 metų intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra didesnė nei 25 metai, skaičius (procentinė dalis pervertintų stebėjimų)** |
|---|---|---|---|
| 1 | 27(44,44%) | 54(68,52%) | 179(49,16%) |
| 2 | 105(45,71%) | 203(48,77%) | 252(49,21%) |
| 3 | 24(25,00%) | 69(53,62%) | 151(48,34%) |

**23 lentelė.** Vilniaus butų pardavimo kainų modeliavimo rezultatai, gauti naudojant kambarių skaičiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (procentinė dalis pervertintų stebėjimų)** |
|---|---|---|---|
| 1 | 36(50,00%) | 94(51,06%) | 139(54,68%) |
| 2 | 48(41,67%) | 162(44,44%) | 360(50,83%) |
| 3 | 2(50,00%) | 54(50,00%) | 196(45,92%) |

Remiantis 22 ir 23 lentelių rezultatais, pagal kiekvieną kambarių skaičiaus reikšmę galima identifikuoti „įdomesnius“ segmentus, kurie pasižymi tiek didesniu prognozuotų stebėjimų (tik tų, kurių prognozės paklaida nebuvo didesnė nei 20%) skaičiumi, tiek didesne procentine dalimi pervertintų stebėjimų:

- Vieno kambario butai pardavimui: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra 6-25 metų intervale (54 stebėjimai, 68,52% pervertintų stebėjimų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 4 km (139 stebėjimai, 54,68% pervertintų stebėjimų);
- Dviejų kambarių butai pardavimui: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 25 metai (252 stebėjimai, 49,21% pervertintų stebėjimų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 4 km (360 įrašų, 54,68% pervertintų stebėjimų);
- Trijų kambarių butai pardavimui: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra 6-25 metų intervale (69 stebėjimai, 53,62% pervertintų stebėjimų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra 1,5-4 km intervale (54 stebėjimai, 50% pervertintų stebėjimų).

Toliau atliktas kainos modeliavimas pagal segmentus, naudojant visų atrinktų butų (nepriklausomai nuo kambarių skaičiaus) pardavimo duomenis (žr. 24 lentelę).

**24 lentelė.** Vilniaus butų pardavimo kainų modeliavimo rezultatai, gauti naudojant pastato amžiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Pastato amžiaus požymio reikšmė** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (procentinė dalis pervertintų stebėjimų)** |
|---|---|---|---|
| 0-5 metų intervale | 4(75,00%) | 31(25,81%) | 121(45,45%) |
| 6-25 metų intervale | 18(66,67%) | 52(44,23%) | 256(53,91%) |
| Didesnė nei 25 metai | 64(37,50%) | 226(51,33%) | 292(49,66%) |

Toliau kainos modeliavimo pagal segmentus procedūra atlikta naudojant Kauno butų pardavimo duomenų rinkinį. Rezultatai pateikiami 25, 26, 27 lentelėse.

**25 lentelė.** Kauno butų pardavimo kainų modeliavimo rezultatai, gauti naudojant kambarių skaičiaus ir pastato amžiaus požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 0-5 metų intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 6-25 metų intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra didesnė nei 25 metai, skaičius (procentinė dalis pervertintų stebėjimų)** |
|---|---|---|---|
| 1 | 4(50,00%) | 1(100%) | 51(43,14%) |
| 2 | 19(63,16%) | 27(33,33%) | 151(51,66%) |
| 3 | 13(69,23%) | 19 (42,11%) | 69(49,28%) |

**26 lentelė.** Kauno butų pardavimo kainų modeliavimo rezultatai, gauti naudojant kambarių skaičiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (procentinė dalis pervertintų stebėjimų)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (procentinė dalis pervertintų stebėjimų)** |
|---|---|---|---|
| 1 | 7(28,57%) | 28(50,00%) | 23(47,83%) |
| 2 | 39(41,03%) | 93(55,91%) | 65(47,69%) |
| 3 | 11(36,36%) | 41(53,66%) | 52(53,85%) |

**27 lentelė.** Kauno butų pardavimo kainų modeliavimo rezultatai, gauti naudojant pastato amžiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Pastato amžiaus požymio reikšmė** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (procentinė** |
|---|---|---|---|

| | (procentinė dalis pervertintų stebėjimų) | (procentinė dalis pervertintų stebėjimų) | dalis pervertintų stebėjimų) |
|---|---|---|---|
| 0-5 metų intervale | 3(33,33%) | 10(80,00%) | 23(60,87%) |
| 6-25 metų intervale | 3(0%) | 19(52,63%) | 25(32,00%) |
| Didesnė nei 25 metai | 51(41,18%) | 131(51,91%) | 89(50,56%) |

## 3.8. Nuomos duomenimis paremtų modelių praktinis taikymas

Panašiai kaip ir su butų pardavimo duomenimis, toliau pateiktas nuomos modelių praktinio taikymo pavyzdys, suteikiant naudingų įžvalgų investuotojams, kurie norėtų įvertinti ilgalaikę perspektyvą – buto atsiperkamumą nuomos pajamų atžvilgiu. Šiame pavyzdyje panaudotas pardavimo-nuomos rodiklis, kuris apibrėžiamas kaip buto faktinė pardavimo kaina, padalinta iš metinių nuomos pajamų. Remiantis šiuo rodikliu, kiekvienam parduodamam butui gautas apytikslis metų skaičius, per kurį jis atsipirktų tik iš nuomos generuojamų pajamų. Šį kartą modeliavimas atliktas treniruojant geriausius modelius su pilnais nuomos duomenų rinkiniais, o testavimui naudoti anksčiau minėti filtruoti butų pardavimo duomenų rinkiniai – tokiu būdu gautos preliminarios nuomos kainos tiems objektams, kurie pirkėjui yra aktualūs tiek gyvenamosios paskirties, tiek atsiperkamumo atžvilgiu. Panašiai kaip ir ankstesniame etape, preliminarioms nuomos kainoms gauti naudotas 5 geriausių regresijos modelių prognozių vidurkis. Toliau pateikiami šio praktinio taikymo pavyzdžio rezultatai, gauti naudojant ankstesniame etape atrinktus butų pardavimo duomenis ir tą pačia segmentavimo strategiją. Kadangi šį kartą nebuvo analizuojamos pardavimo kainų prognozės ir jų tikslumas, panaudoti visų, globalius apribojimus atitinkančių alternatyvų duomenys (žr. 28, 29, 30 lenteles).

**28 lentelė.** Vilniaus butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei kambarių skaičiaus ir pastato amžiaus požymių pagalba sukurtus segmentus

| Kambarių skaičius | Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 0-5 metų intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) | Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 6-25 metų intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) | Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra didesnė nei 25 metai, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) |
|---|---|---|---|
| 1 | 49(19,72) | 67(19,70) | 295(15,98) |
| 2 | 153(19,55) | 240(21,39) | 369(18,37) |
| 3 | 27(18,06) | 76(20,86) | 210(18,88) |

**29 lentelė.** Vilniaus butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei kambarių skaičiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| Kambarių skaičius | Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) | Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) | Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais) |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | 66(18,67) | 180(16,34) | 177(17,17) |
| 2 | 62(20,11) | 275(19,35) | 436(19,59) |
| 3 | 2(16,97) | 80(19,18) | 239(19,30) |

**30 lentelė.** Vilniaus butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei pastato amžiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Pastato amžiaus požymio reikšmė** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** |
|---|---|---|---|
| 0-5 metų intervale | 7(20,05) | 72(20,17) | 150(19,02) |
| 6-25 metų intervale | 21(19,07) | 80(21,70) | 282(20,93) |
| Didesnė nei 25 metai | 101(19,41) | 381(17,25) | 392(17,66) |

Remiantis 28 ir 29 lentelių rezultatais, galima identifikuoti Vilniaus butų nuomos rinkos segmentus, kurie investuotojui yra įdomūs dėl pakankamai didelio alternatyvų skaičiaus ir kiek įmanoma mažesnės vidutinės pardavimo-nuomos kainų santykio reikšmės:

- Vieno kambario butai nuomai: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 25 metai (295 stebėjimai, pardavimo-nuomos santykio vidurkis 15,96 metų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra 1,5-4 km intervale (180 stebėjimų, pardavimo-nuomos santykio vidurkis 17,17 metų);
- Dviejų kambarių butai nuomai: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 25 metai (369 stebėjimai, pardavimo-nuomos santykio vidurkis 18,37 metų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra 1,5-4 km intervale (275 stebėjimai, pardavimo-nuomos santykio vidurkis 19,35 metų);
- Trijų kambarių butai nuomai: pagal pastato amžiaus požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 25 metai (210 stebėjimų, pardavimo-nuomos santykio vidurkis 18,88 metų), o pagal „ats_centras“ požymį išsiskiria segmentas, kurio visų stebėjimų reikšmės yra didesnės nei 4 km (239 stebėjimai, pardavimo-nuomossantykio vidurkis 19,3 metų).

Ta pati procedūra pakartota naudojant Kauno butų pardavimo duomenis bei nuomos kainos prognozavimo modelius, rezultatai pateikti 31, 32, 33 lentelėse.

**31 lentelė.** Kauno butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei kambarių skaičiaus ir pastato amžiaus požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 0-5 metų intervale, skaičius (pardavimo kainos ir** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra 6-25 metų intervale, skaičius (pardavimo kainos ir** | **Bendras prognozuotų stebėjimų, kurių pastato amžiaus požymio reikšmė yra didesnė nei 25 metai, skaičius (pardavimo kainos ir** |
|---|---|---|---|

| | **nuomos kainos santykio vidurkis, metais)** | **nuomos kainos santykio vidurkis, metais)** | **nuomos kainos santykio vidurkis, metais)** |
|---|---|---|---|
| 1 | 4(20,21) | 1(15,46) | 82(13,89) |
| 2 | 28(19,55) | 33(21,18) | 213(17,38) |
| 3 | 14(19,23) | 22(20,41) | 115(18,24) |

**32 lentelė.** Kauno butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei kambarių skaičiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Kambarių skaičius** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** |
|---|---|---|---|
| 1 | 12(17,71) | 44(14,45) | 33(12,70) |
| 2 | 59(21,27) | 137(17,24) | 79(17,02) |
| 3 | 23(18,58) | 64(18,59) | 67(18,57) |

**33 lentelė.** Kauno butų nuomos kainų modeliavimo rezultatai, gauti naudojant pardavimo duomenis bei pastato amžiaus ir „ats_centras“ požymių pagalba sukurtus segmentus

| **Pastato amžiaus požymio reikšmė** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 0-1,5 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra 1,5-4 km intervale, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** | **Bendras prognozuotų stebėjimų, kurių „ats_centras“ požymio reikšmė yra didesnė nei 4 km, skaičius (pardavimo kainos ir nuomos kainos santykio vidurkis, metais)** |
|---|---|---|---|
| 0-5 metų intervale | 7(20,44) | 12(18,78) | 27(19,60) |
| 6-25 metų intervale | 3(23,24) | 26(21,15) | 27(20,13) |
| Didesnė nei 25 metai | 84(20,02) | 204(16,51) | 122(15,49) |

## Išvados

1. Atlikus Lietuvos ir užsienio mokslinės literatūros analizę, išsiaiškinta, kad būsto kainų modeliavimas yra kompleksinis uždavinys, kuriam spręsti taikoma daug skirtingų metodų. Pastebėta, kad tyrimų autoriai dažniausiai šią problemą sprendžia kaip laiko eilučių arba kaip regresijos uždavinį. Surasta ir tokių tyrimų, kuriuose integruojami kelių skirtingų klasių metodai, pavyzdžiui, turint ilgesnio laikotarpio duomenis panaudojami ir makroekonominių, demografinių rodiklių pokyčiai arba absoliutinės reikšmės, tada modeliavimas atliekamas pagal „slenkančio lango“ duomenų išskaidymo strategiją. Vis dėlto, atsižvelgiant į turimų duomenų apimtį ir išsikeltus tikslus, pasirinkta apsiriboti regresijos klasės metodais. Perėjus prie regresijos modelių analizės, pastebėta, jog vis dažniau atsisakoma klasikinių modelių ir pereinama prie pažangesnių mašininio mokymosi algoritmų. Taip pat pastebėta, kad ypatingai literatūroje išsiskyria *Random Forest* ir *XGBoost* modeliai, kurie pasižymi gebėjimu įvertinti netiesinius ryšius tarp būsto kainos ir paaiškinamųjų požymių. Daugelyje nagrinėtų tyrimų šie modeliai pasirodė kaip vieni tiksliausių būsto kainos modeliavimo įrankių, o jų papildomas pranašumas yra paprastas hiperparametrų optimizavimo procesas, kuris nereikalauja daug resursų;
2. Naudojant *Selenium* ir *BeautifulSoup* bibliotekas, buvo sėkmingai surinkti daugiau nei 5000 Vilniaus ir Kauno miestuose parduodamų ir nuomojamų butų duomenys. Automatizuoti įrankiai leido išsirinkti standartiniuose *HTML* failuose esančius elementus, taip pat, naudojant logines funkcijas, panaudota skelbimų aprašymuose esanti informacija bei „Google Maps“ svetainėje esančios objekto koordinatės. Nors buvo svarstyta į modeliavimo etapą įtraukti ir laiko požymius (pvz. skelbimo įkėlimo datą), galima teigti, kad dėl trumpo duomenų rinkimo laikotarpio ir nuolat skelbimų autorių koreguojamų kainų, šie požymiai nebūtų suteikę daug naudos;
3. Atlikus surinktų duomenų žvalgomąją analizę, iš pirminių 26 paaiškinamųjų požymių buvo suformuoti nauji požymiai, kurie padėjo išgauti papildomą naudingą informaciją, o koreliacinės analizės rezultatai padėjo įsivertinti, kurių požymių sąveikas vertėtų išbandyti modeliavimo etape. Toliau iš duomenų rinkinių buvo pašalintos išskirtys pagal buto pardavimo arba nuomos kainą, taip pat pagal plotą ir kambarių skaičių, naudojant artimiausių kaimynų metodą, užpildytos trūkstamos reikšmės. Visos šios duomenų paruošimo procedūros sudarė galimybes kurti tikslesnius regresijos modelius ir tokiu būdu gauti patikimesnes prognozes;
4. Atlikus modeliavimo ir tikslumo vertinimo procedūras, paaiškėjo, kad skirtingos požymių atrankos strategijos turėjo reikšmingą įtaką modelių tikslumui. Klasikiniai regresijos modeliai geriausiai veikė naudojant reguliarizuotų regresijų atrankos pagalba gautus požymius, tuo tarpu sprendimų medžių principu veikiantys modeliai pasiekė didžiausią tikslumą naudojant pilnus paaiškinamųjų požymių rinkinius. Pastebėta, kad hiperparametrų optimizavimas, naudojant *Optuna* biblioteką, padėjo pagerinti beveik visų ištestuotų modelių rezultatus. Taip pat pastebėta, kad visuose modeliavimo etapuose medžių modeliai pranoko klasikinius regresijos modelius;
5. Atlikus butų pardavimo kainos modeliavimą buvo identifikuoti didžiausiu tikslumu pasižymintys modeliai abiems miestams – Vilniaus butų pardavimo duomenų atveju geriausius rezultatus, pagal *MAPE* rodiklį, pademonstravo *XGBoostRegressor* modelis (11,3%), o naudojant Kauno butų pardavimo duomenis, geriausias rezultatas gautas pasitelkus *GradientBoostingRegressor* modelį (15,54%). Remiantis šių modelių sugeneruotais paaiškinamųjų požymių svarbumo koeficientais, buvo nustatyta, kad abiejų miestų atveju daugiausia įtakos kvadratinio metro pardavimo kainos kintamumui turėjo pastato amžiaus ir Euklido atstumo iki miesto centro požymiai. Šie požymiai, kartu su buto kambarių skaičiumi, buvo naudojami atliekant modeliavimo procedūras pagal atskirus rinkos segmentus.

## Rekomendacijos

1. Duomenų analitikui, siekiančiam kuo tiksliau prognozuoti būsto kainas, rekomenduojama plėsti duomenų šaltinių įvairovę. Šio tyrimo metu pastebėta, kad modelių prognozavimo tikslumas yra glaudžiai susijęs su turimų stebėjimų skaičiumi, todėl tolimesniuose tyrimuose, be „aruodas.lt“ svetainės, būtų naudinga integruoti ir kitų nekilnojamojo turto skelbimų portalų duomenis. Be to, įtraukiant vaizdinę informaciją (pvz. vektorizuotą skelbimo nuotraukų informaciją) ar papildomus semantinius bei kontekstinius duomenis (pvz. papildomus atstumų požymius iš „Google Maps“) galima padidinti paaiškinamųjų požymių skaičių, tačiau reikėtų atkreipti dėmesį į tai, kad šis papildomų požymių kūrimas turėtų būti derinamas atsižvelgiant į turimų stebėjimų skaičių;
2. Būsto pirkėjams rekomenduojama remtis ne tik modelių prognozuojamomis kainomis, bet ir atkreipti dėmesį į tai, kuriuose rinkos segmentuose regresijos modeliai geriausiai paaiškina kainų elgseną. Tyrime atlikto modeliavimo pagal segmentus rezultatai parodė, kad daugelyje grupių pervertintų ir visų tos grupės stebėjimų santykis yra artimas 50%, tačiau kai kuriose grupėse (pvz. Vilniaus pardavimo duomenų imtyje, susidedančioje tik iš 1 kambario butų, kurių pastato amžiaus požymio reikšmė yra 6-25 metų intervale) pastebėti reikšmingi nuokrypiai. Rekomenduojama giliau tirti tuos segmentus, kuriuose modeliai sistemingai pervertina arba nuvertina butų kainas ir tokiu būdu identifikuoti tuos veiksnius, kurie yra nepastebimi naudojant tik paaiškinamųjų požymių informaciją.

## Literatūros sąrašas


1. Tostevin, P., & Rushton, C. (2022). The total value of global real estate : Property remains the world‘s biggest store of wealth. *Savills Impacts* [žiūrėta 2025-05-16]. Prieiga per internetą: https://impacts.savills.com/market-trends/the-total-value-of-global-real-estate-property-remains-the-worlds-biggest-store-of-wealth.html
2. Maclennnan, D., Leishman, C., & Goel, S. (2023). How does the housing market affect financial and economic stability? *Economics Observatory* [žiūrėta 2025-05-16]. Prieiga per internetą: https://www.economicsobservatory.com/how-does-the-housing-market-affect-financial-and-economic-stability
3. Farzanegan, M. R., & Gholipour, H. F. (2024). How might house prices affect workers productivity in OECD economies? *Economics Observatory* [žiūrėta 2025-05-16]. Prieiga per internetą: https://www.economicsobservatory.com/how-might-house-prices-affect-workers-productivity-in-oecd-economies
4. Schwartz, D. (2016). The Importance of Affordable Housing to Economic Competitiveness. *Economic Development Journal*, 15, 1, p. 40-4 [žiūrėta 2025-05-16]. Prieiga per internetą: https://www.iedconline.org/clientuploads/Economic%20Development%20Journal/EDJ_16_Winter_Schwartz.pdf
5. *Lietuvos statistikos departamentas*. (2025). Statistinių rodiklių duomenų bazė, būsto kainų indeksai (2015 m. = 100) [žiūrėta 2025-03-02]. Prieiga per internetą: https://osp.stat.gov.lt/statistiniu-rodikliu-analize#/
6. Tripathi, S. (2019). Macroeconomic determinants of housing prices: A cross country level analysis (MPRA Paper No. 98089). Munich: Personal RePEc Archive [žiūrėta 2025-03-05]. Prieiga per internetą: https://mpra.ub.uni-muenchen.de/98089/
7. Tupėnaitė, L., Kanapeckienė, L., Naimavičienė, J. (2017). Determinants of housing market fluctuations: Case study of Lithuania. *Procedia Engineering*, 172, p. 1169-1175 [žiūrėta 2025-03-06]. DOI: https://doi.org/10.1016/j.proeng.2017.02.136
8. Naruševičius, L., Ramanauskas, T., Gudauskaitė, L., Reichenbachas, T. (2019). Lithuanian house price index: Modelling and forecasting (Occasional Paper Series No. 28). Vilnius: Lietuvos Bankas [žiūrėta 2025-03-06]. Prieiga per internetą: http://hdl.handle.net/11159/3580
9. Jadevičius, A., & Parsa, A. (2014). An empirical analysis of real estate cycles in the Lithuanian housing market. *Journal of Real Estate Literature*, 22, no. 1, p. 69-81 [žiūrėta 2025-05-16]. Prieiga per internetą: https://www.researchgate.net/publication/265612206_An_Empirical_Analysis_of_Real_Estate_Cycles_in_the_Lithuanian_Housing_Market
10. Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic modelling*, 45, p. 259-267 [žiūrėta 2025-04-04]. DOI: https://doi.org/10.1016/j.econmod.2014.10.050
11. Usman, H., Lizam, M., & Burhan, B. (2020). Review of issues in the conventional hedonic property pricing model. *In proceedings of the 2nd African International Conference on Industrial Engineering and Operations Management* [interaktyvus] p. 2806-2816. IEOM Society International [žiūrėta 2025-04-04]. Prieiga per internetą: https://www.ieomsociety.org/harare2020/papers/631.pdf
12. Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., & Guo, J. (2017). Exploring determinants of housing prices in Beijing: an enhanced hedonic regression with open access POI data. *ISPRS international*

*journal of geo-information*, 6(11), staipsnis 358 [žiūrėta 2025-04-05]. DOI: https://doi.org/10.3390/ijgi6110358

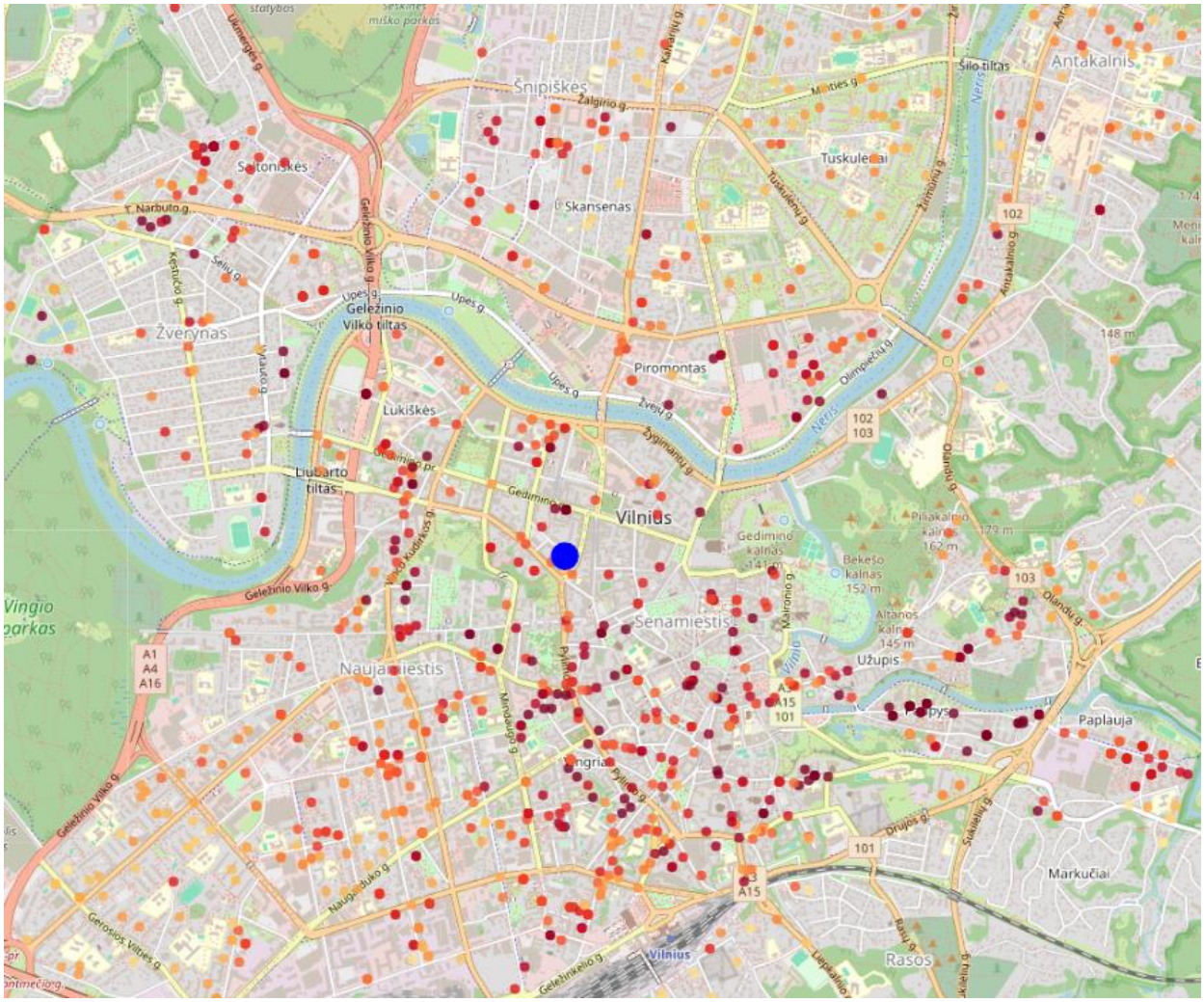
13. Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1), p. 34-55 [žiūrėta 2025-04-06]. DOI: https://doi.org/10.2307/3146991
14. Cropper, M. L., & Mcconnell, K. (1988). On the choice of functional form for hedonic price functions. *Review of economics and statistics*, 70(4), p. 668-675 [žiūrėta 2025-04-07]. DOI: https://doi.org/10.1016/S0095-0696(02)00013-X
15. Choi, S., & Yi. M. Y. (2021). Computational valuation model on housing price using pseudo self comparison method. *Sustainability*, 13(20), straipsnis 11489 [žiūrėta 2025-04-07]. DOI: https://doi.org/10.3390/su132011489
16. Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management,* 24(3), p. 140-152 [žiūrėta 2025-04-08]. DOI: https://doi.org/10.3846/ijspm.2020.11544
17. Antipov, E., & Pokryshevskaya, E. (2012). Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. *SSRN* [žiūrėta 2025-04-08]. Prieiga per internetą: https://ssrn.com/abstract=1729653
18. Teang, K., & Lu, Y. (2021). Property valuations by machine learning and hedonic pricing models: A case study on Swedish residential property. Magistro darbas, KTH Royal Institute of Technology [žiūrėta 2025-04-09]. Priega per internetą: https://www.diva-portal.org/smash/get/diva2:1576509/FULLTEXT01.pdf
19. Truong, Q., & Nguyen, M. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, p. 433-442 [žiūrėta 2025-04-09]. DOI: https://doi.org/10.1016/j.procs.2020.06.111
20. Kmen, C. Navratil, G., & Giannopoulos, I. (2024). Location, location, location: The power of neighborhoods for apartment price predictions based on transaction data. *ISPRS International Journal of Geo-Information*, 13(12), straipsnis 425 [žiūrėta 2025-04-10]. DOI: https://doi.org/10.3390/ijgi13120425
21. Kalliola, J., Dzikienė, J., K., & Damaševičius, R. (2021). Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ Computer Science*, [žiūrėta 2025-04-10]. DOI: https://doi.org/10.7717/peerj-cs.444
22. Deaconu, A., Buiga, A., & Tothazan, H. (2022). Real estate valuation models performance in price prediction. International *Journal of Strategic Property Management*, 26(2), p. 86-105 [žiūreta 2025-04-11]. DOI: https://doi.org/10.3846/ijspm.2022.15962
23. Potrawa, T., & Tetereva, A. (2022). How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, 144, p. 50-65 [žiūrėta 2025-04-11]. DOI: https://doi.org/10.1016/j.jbusres.2022.01.027
24. Caroni, C. M. (2023). Rental price prediction using machine learning: A French case-study. Magistro darbas, Politecnico di Milano [žiūrėta 2025-04-12]. Priega per internetą: https://www.politesi.polimi.it/handle/10589/203673
25. Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics*, 3(1), p. 30-45 [žiūrėta 2025-04-12]. DOI: https://doi.org/10.3390/analytics3010003

26. Adomavičius, S. (2022). Nekilnojamojo turto vertės nustatymas pasitelkiant mašininio mokymosi technikas. Magistro darbas, Kauno Technologijos Universitetas, Matematikos ir gamtos mokslų fakultetas [žiūrėta 2025-04-10]. Prieiga per internetą: https://epubl.ktu.edu/object/elaba:132442661/132442661.pdf
27. *Similarweb*. Aruodas.lt – Traffic and Engagement Overview. (2025). [žiūrėta 2025-04-23]. Prieiga per internetą: https://www.similarweb.com/website/aruodas.lt/
28. Zhang, Q. (2021). Housing Price Predicition Based on Multiple linear Regression. *Scientific Programming*, straipnis 7678931 [žiūrėta 2025-04-25]. DOI: https://doi.org/10.1155/2021/7678931
29. Preethi, S., Murthy, D. H. R., Hiremani, V., Raghavendra, M. D., & Sapna, R. (2025). Optimizing Polynomial and Regularization Techniques for Enhanced Housing Price Prediction Accuracy. *SN Computer Science*, 6(1), straipsnis 96 [žiūrėta 2025-04-25]. DOI: https://doi.org/10.1007/s42979-024-03578-7
30. Kulkarni, N. (2022). How to build a Bayesian Ridge Regression Model with Full Hyperparameter integration. *Medium* [žiūrėta 2025-04-26]. Prieiga per internetą: https://medium.com/data-science/how-to-build-a-bayesian-ridge-regression-model-with-full-hyperparameter-integration-f4ac2bdaf329
31. Ariyanti, N. P., Triayudi, A., & Sari, R. T. K. (2024). Analysis of K-NN Algorithm and Linear Regression to Predict House Prices in Jabodetabek. *SaNa: Journal of Blockchain, NFTs and Metaverse Technology*, 2, p. 65-71 [žiūrėta 2025-04-26]. DOI: https://doi.org/10.58905/sana.v2i1.265
32. Fernandez, J. M. (2020). Tree ensembles: theory and practice. *Towards Data Science* [žiūrėta 2025-04-26]. Prieiga per internetą: https://towardsdatascience.com/tree-ensembles-theory-and-practice-1cf9eb27781/
33. Valverde, J. (2019). Metrics evaluation: MSE, RMSE, MAE, and MAPE. *Medium* [žiūrėta 2025-04-26]. Prieiga per internetą: https://medium.com/@jonatasv/metrics-evaluation-mse-rmse-mae-and-mape-317cab85a26b
34. Elmuna, E. A. F., Chamidy, T., & Nugroho, F. (2023). Optimization of the random forest method using principal component analysis to predict house prices: A case study of house prices in Malang City. *International Journal of Advances in Data and Information Systems*, 4(2), p. 155-166 [žiūrėta 2025-04-27]. DOI: https://doi.org/10.25008/ijadis.v4i2.1290
35. Mazraeh, A. (2023). Polynomial features: A comprehensive guide from basics to advanced. *Medium* [žiūrėta 2025-04-28]. Prieiga per internetą: https://medium.com/@adnan.mazraeh1993/polynomial-features-a-comprehensive-guide-from-basics-to-advanced-5f18c430a137
36. Brownlee, J. (2020). Power transforms with Scikit-Learnn. *Machine Learning Mastery* [žiūrėta 2025-05-01]. Prieiga per internetą: https://machinelearningmastery.com/power-transforms-with-scikit-learn/
37. Malato, G. (2021). Feature selection in machine learning using Lasso regression. *Your Data Teacher* [žiūrėta 2025-05-01]. Prieiga per internetą: https://www.yourdatateacher.com/2021/05/05/feature-selection-in-machine-learning-using-lasso-regression/

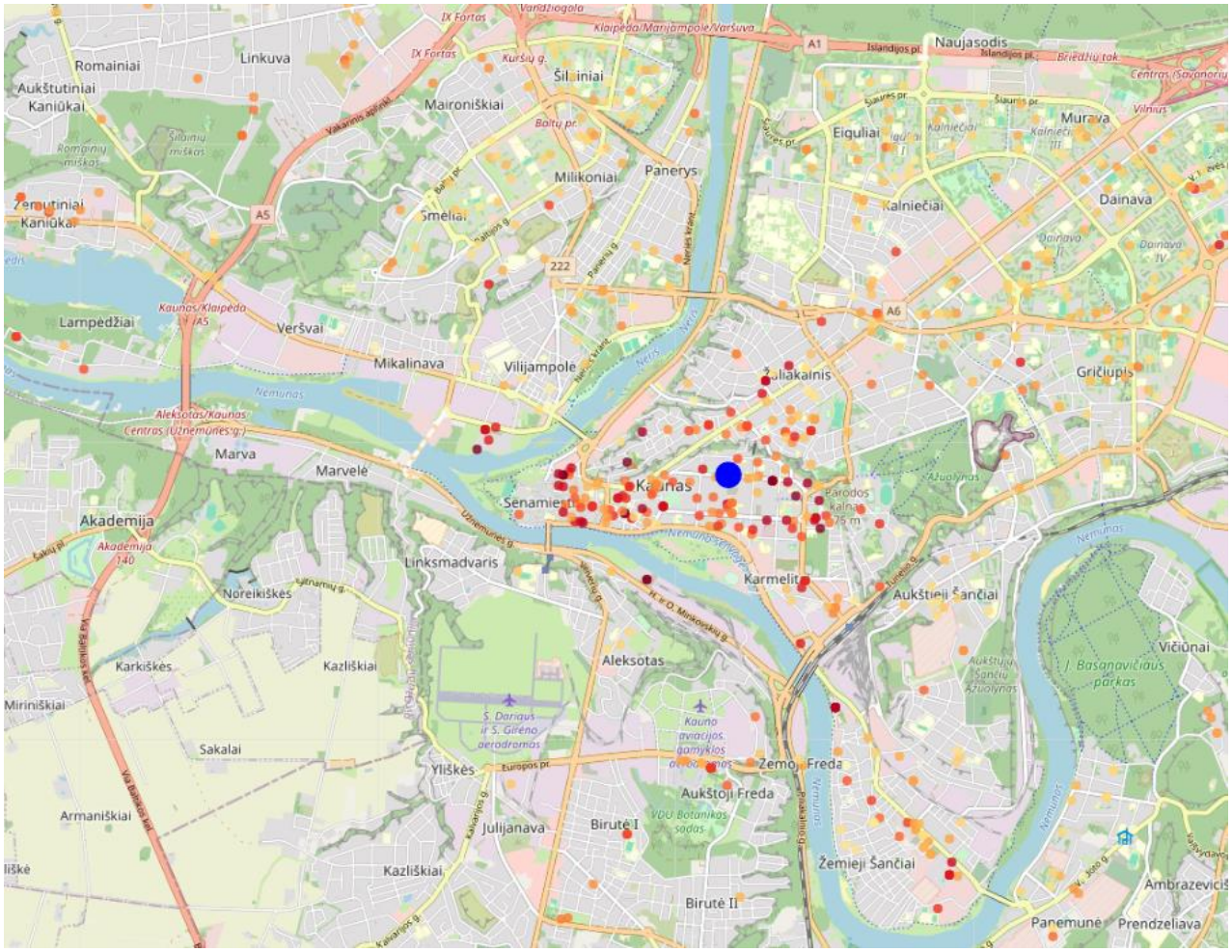
38. Vemula, S. (2022). Boruta feature selection explained in Python. *Medium* [žiūrėta 2025-05-01]. Prieiga per internetą: https://medium.com/geekculture/boruta-feature-selection-explained-in-python-7ae8bf4aa1e7
39. Kee, T., & Ho, W. K. O. (2024). Optimizing machine learning models for urban sciences: a comparative analysis of hyperparameter tuning methods. *Preprints* [žiūrėta 2025-05-02]. DOI: https://doi.org/10.20944/preprints202406.0264.v2

## Priedai

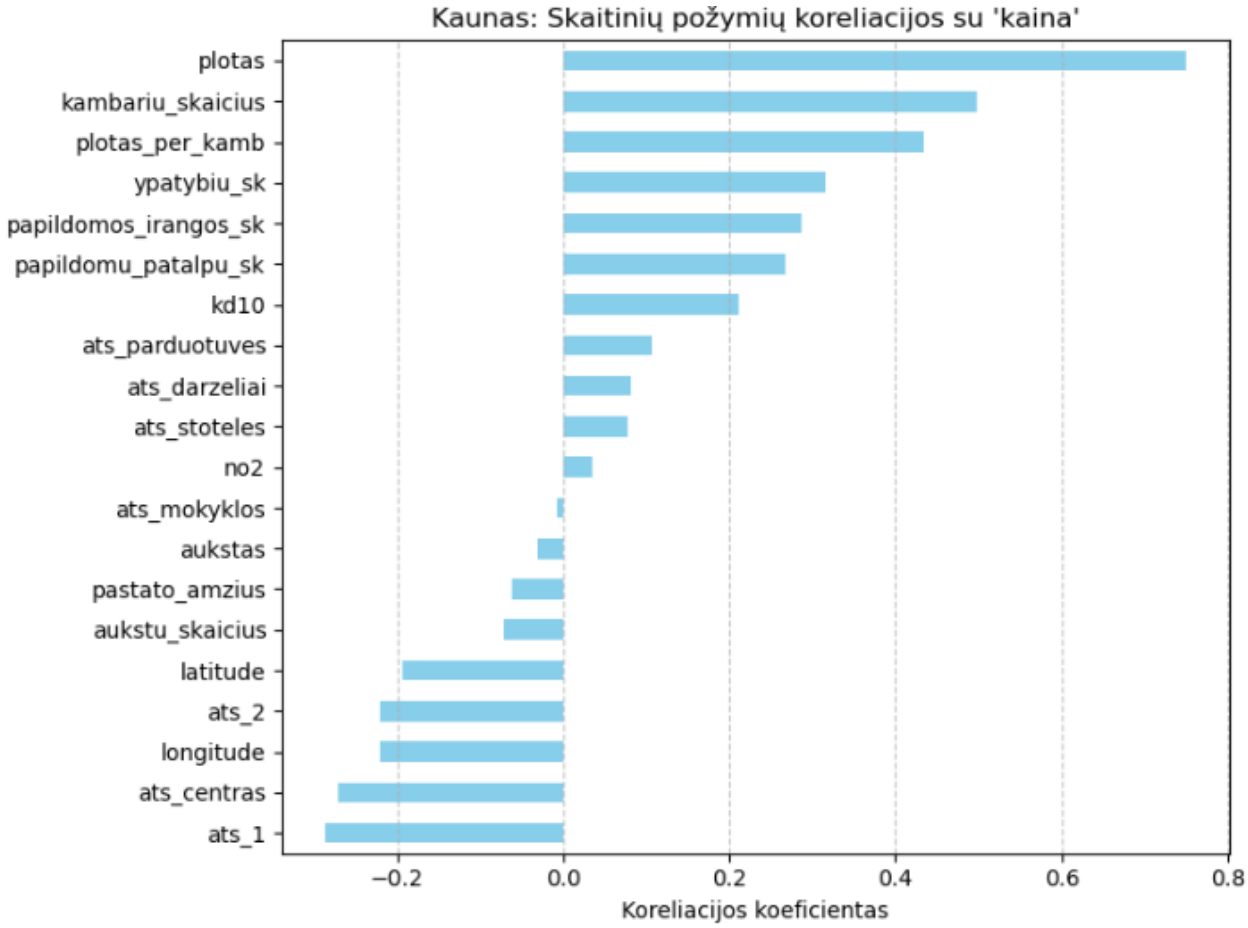
**1 priedas. Vilniaus butų kv. metro pardavimo kainų žemėlapis su „centriniu“ tašku, gautu iš ilgumos ir platumos požymių, naudojant *KernelDensity* funkciją**



**2 priedas. Kauno butų kv. metro pardavimo kainų žemėlapis su „centriniu“ tašku, gautu iš ilgumos ir platumos požymių, naudojant *KernelDensity* funkciją**



**3 priedas. Kauno butų pardavimo duomenų rinkinio skaitinių požymių ir buto kainos tarpusavio koreliacijos**



## 4 priedas. Visų modelių hiperparametrų optimizavimo rezultatai, gauti naudojant Vilniaus butų nuomos duomenų rinkinį

| **Modelis** | **Geriausias požymių rinkinys** | **Ar logaritmuoti „kv_kaina“ ?** | **Patikrintų hiperparametrų kombinacijų skaičius** | **Bendras skaičiavimo laikas** | **Geriausia MAPE reikšmė (%)** |
|---|---|---|---|---|---|
| LinearRegression | Reg. atrankos | Taip | - | - | 15,54 |
| Ridge | Originalus | Taip | 100 | 6,6 s | 15,37 |
| Lasso | Originalus | Taip | 100 | 6,9 s | 15,33 |
| ElasticNet | Originalus | Taip | 100 | 7,6 s | 15,32 |
| BayesianRidge | Originalus | Taip | 100 | 13,3 s | 15,38 |
| SGD | Originalus | Ne | 100 | 23,7 s | 16,55 |
| SVR | Reg. atrankos | Taip | 100 | 50 s | 14,36 |
| RandomForest | Originalus | Ne | 100 | 3 min 23 s | 14,07 |
| GradientBoostingRegressor | Originalus | Ne | 100 | 17 min 31 s | **13,55** |
| BaggingRegressor | Mišrus | Ne | 100 | 3 min 47 s | 14,15 |
| ExtraTreesRegressor | Boruta atrankos | Ne | 100 | 1 min 36 s | 14,73 |
| XGBoostRegressor | Mišrus | Ne | 100 | 2 min 2 s | 13,72 |
| LGBMRegressor | Originalus | Ne | 100 | 4 min 54 s | 14,45 |
| HistGradientBoostingRegressor | Originalus | Ne | 100 | 9 min 57 s | 14,56 |
| AdaBoostRegressor | Mišrus | Ne | 100 | 9 min 28 s | 16,85 |
| KNeighborsRegressor | Reg. atrankos | Taip | 100 | 6,9 s | 14,41 |

## 5 priedas. Visų modelių hiperparametrų optimizavimo rezultatai, gauti naudojant Kauno butų nuomos duomenų rinkinį

| **Modelis** | **Geriausias požymių rinkinys** | **Ar logaritmuoti „kv_kaina“ ?** | **Patikrintų hiperparametrų kombinacijų skaičius** | **Bendras skaičiavimo laikas** | **Geriausia MAPE reikšmė (%)** |
|---|---|---|---|---|---|
| LinearRegression | Mišrus | Taip | - | - | 19,05 |
| Ridge | Originalus | Taip | 100 | 2,1 s | 18,85 |
| Lasso | Originalus | Taip | 100 | 2,2 s | 18,92 |
| ElasticNet | Originalus | Taip | 100 | 2,7 s | 18,84 |
| BayesianRidge | Originalus | Taip | 100 | 6,2 s | 18,86 |
| SGD | Originalus | Taip | 100 | 4,1 s | 19,72 |
| SVR | Originalus | Taip | 100 | 8,6 s | 19,72 |
| RandomForest | Mišrus | Ne | 100 | 3 min 41 s | 17,17 |
| GradientBoostingRegressor | Originalus | Ne | 100 | 4 min 44 s | 16,75 |
| BaggingRegressor | Mišrus | Ne | 100 | 2 min 12 s | 17,09 |
| ExtraTreesRegressor | Boruta atrankos | Ne | 100 | 1 min 48 s | 17,11 |
| CatBoostRegressor | Mišrus | Ne | 100 | 22 min 18 s | **16,33** |
| XGBoostRegressor | Originalus | Ne | 100 | 1 min 59 s | 16,57 |
| LGBMRegressor | Originalus | Ne | 100 | 1 min 29 s | 18,01 |
| HistGradientBoostingRegressor | Mišrus | Ne | 100 | 5 min 19 s | 17,77 |
| AdaBoostRegressor | Originalus | Taip | 100 | 5 min 31 s | 18,24 |