****

**Kauno technologijos universitetas**

Informatikos Fakultetas

**Input & output analysis, Decision tree, Random forest, KNN** Laboratorinis darbas 1

|  |  |
| --- | --- |
|  |  |
| **Ignas Biekša**  Studentas, KTU El. p.: ignas.bieksa@ktu.edu | (parašas) (data) |
|  |  |

Vilnius, 2024

Turinys

[1. Problema 3](#_Toc178698047)

[1.1. Problemos sprendimui naudoti įrankiai 3](#_Toc178698048)

[2. Duomenų analizė ir paruošimas 4](#_Toc178698049)

[2.1. Duomenų tipų ir ypatybių nustatymas 4](#_Toc178698050)

[2.2. Duomenų kokybės ataskaita kiekvienai ypatybei 4](#_Toc178698051)

[2.3. Distribution Characteristics 6](#_Toc178698052)

[2.4. Išvestinės savybės 10](#_Toc178698053)

[2.5. Duomenų paruošimas 10](#_Toc178698054)

[3. KNN, DT, RF ir ANN implementacijos, geriausios hiperparametrų aibės ir ataskaita 11](#_Toc178698055)

[3.1. KNN ataskaita, geriausi hiperparametrai 11](#_Toc178698056)

[3.2. DT ataskaita, geriausi hiperparametrai 12](#_Toc178698057)

[3.3. RF ataskaita, geriausi hiperparametrai 14](#_Toc178698058)

[3.4. ANN ataskaita, geriausi hiperparametrai 15](#_Toc178698059)

[4. Apibendrinimas ir išvados 17](#_Toc178698060)

[Šaltiniai 18](#_Toc178698061)

# Problema

Based on the given data of historical real estate transactions create the decision-making model (DMM) which aims to predict prices of new real estate objects.

## Problemos sprendimui naudoti įrankiai

Problemai spręsti buvo naudojama Python 3.12 virtuali aplinka, bei bibliotekos:

* sklearn – skirtingų modelių implementacijoms, bei metrikų skaičiavimui;
* matplotlib – diagramų generavimui;
* pandas – duomenų tvarkymui;
* numpy – papildomiems aritmetiniams veiksmams;
* ucimlrepo – duomenų rinkinio atsisiuntimui.

# Duomenų analizė ir paruošimas

## Duomenų tipų ir ypatybių nustatymas

Žemiau pateikta lentelė su kiekvienos duomenų rinkinio stulpelio ypatybėmis ir tipais.

**1 lentelė.** Duomenų tipai ir ypatybės.

| Ypatybės pavadinimas | Tipas | Ypatybė |
| --- | --- | --- |
| index | Ordinal | Categorical |
| category | Categorical | Categorical |
| title | Textual | Categorical |
| body | Textual | Categorical |
| amenities | Categorical | Categorical |
| bathrooms | Numeric | Continuous |
| bedrooms | Numeric | Continuous |
| currency | Categorical | Categorical |
| fee | Categorical | Categorical |
| has\_photo | Categorical | Categorical |
| pets\_allowed | Categorical | Categorical |
| price | Numeric | Continuous |
| price\_display | Textual | Categorical |
| price\_type | Categorical | Categorical |
| square\_feet | Numeric | Continuous |
| address | Textual | Categorical |
| cityname | Categorical | Categorical |
| state | Categorical | Categorical |
| latitude | Numeric | Continuous |
| longitude | Numeric | Continuous |
| source | Categorical | Categorical |
| time | Interval | Continuous |

### 

## Duomenų kokybės ataskaita kiekvienai ypatybei

Sekančiame duomenų apdorojimo žingsnyje buvo įvertinta skirtingų tipų duomenų kokybė. Žemiau pateiktos ištisinių ir kategoriškų duomenų lentelės.

**2 lentelė.** Ištisinių (angl. continuous) duomenų kokybės lentelė.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Missing rows %** | | **Cardinality** | **Min** | | **Q1** | | **Mean** |
| bathrooms | 99757 | 0,069 | | 16 | 1 | | 1 | | 1,445 |
| bedrooms | 99696 | 0,130 | | 10 | 0 | | 1 | | 1,728 |
| price | 99821 | 0,005 | | 3691 | 100 | | 1014 | | 1527,208 |
| square\_feet | 99821 | 0,005 | | 2540 | 37,625 | | 729 | | 956,402 |
| latitude | 99797 | 0,029 | | 7215 | 19,574 | | 33,7496 | | 31476,040 |
| longitude | 99795 | 0,031 | | 7272 | -159,370 | | -104,792 | | -91,562 |
| time | 99820 | 0,006 | | 75541 | 1544174418 | | 1550831773 | | 1559689694 |
| **Feature** | **Median** | | **Q3** | | | **Max** | | **Standard Deviation** | |
| bathrooms | 1 | | 2 | | | 9 | | 0,547 | |
| bedrooms | 2 | | 2 | | | 9 | | 0,749 | |
| price | 1350 | | 1795 | | | 52500 | | 903,636 | |
| square\_feet | 900 | | 1115 | | | 50000 | | 417,261 | |
| latitude | 37,230 | | 39,954 | | | 1568782042 | | 7022820,001 | |
| longitude | -84,562 | | -77,585 | | | -68,779 | | 15,819 | |
| time | 1568745092 | | 1568767125 | | | 1577391425 | | 11044865,586 | |

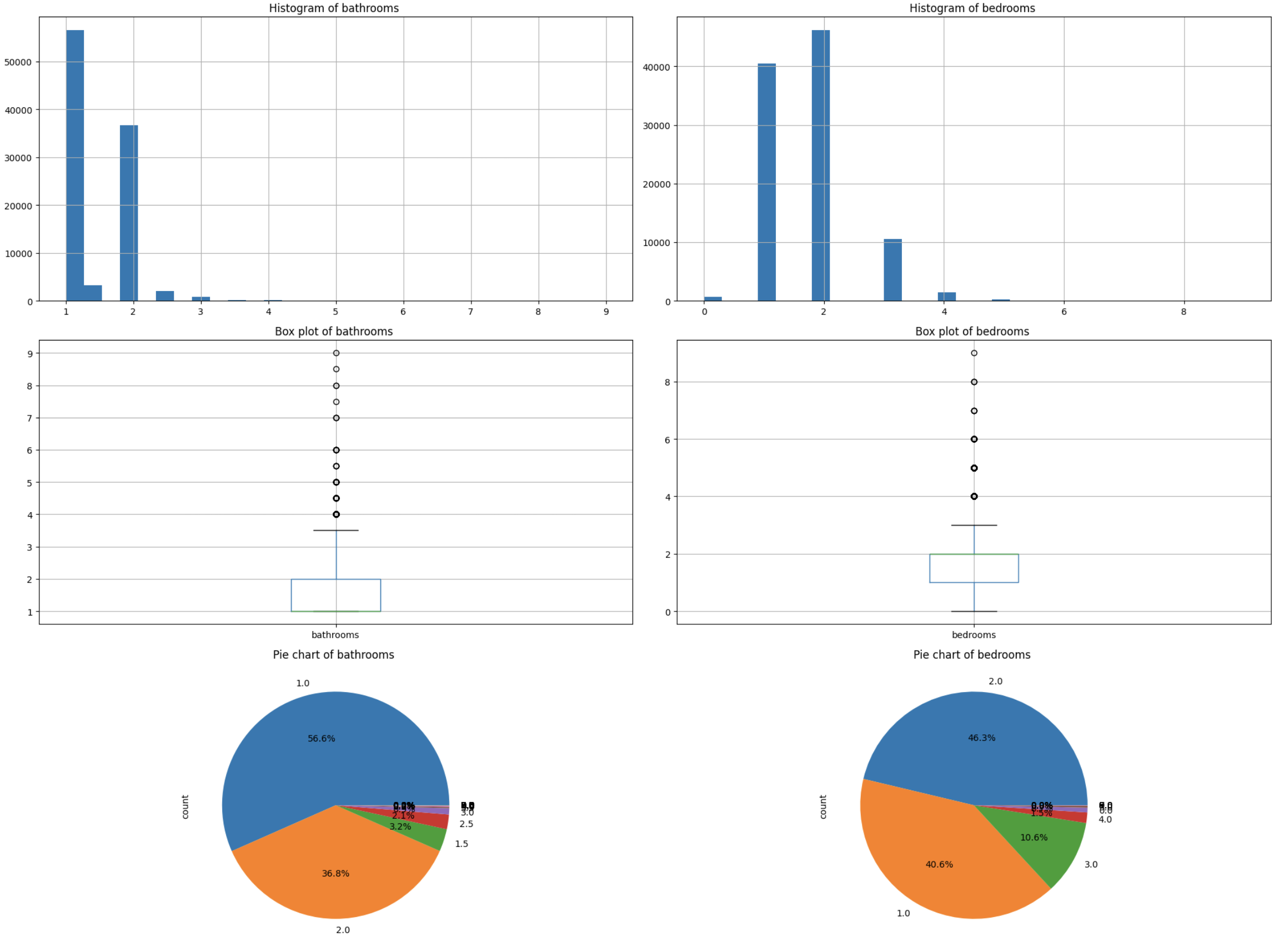
Antroje lentelėje buvo pastebėta, kad ne visi įrašai turi nurodytą kainą. Kadangi sprendžiamas uždavinys yra kainos prognozavimas, įrašai be kainos buvo panaikinti. **3 lentelė.** Kategoriškų duomenų kokybės lentelė.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | | **Count** | | **Cardinality** | | **Top** | | **Top Freq.** | | **% Missing rows** | |
| category | | 99826 | | 10 | | housing/  rent/apartment | | 99762 | | 0 | |
| title | | 99826 | | 58660 | | Apartment in great location | | 1066 | | 0 | |
| body | | 99826 | | 94806 | | When searching for a pet-friendly... | | 45 | | 0 | |
| amenities | | 83749 | | 9843 | | Parking | | 6204 | | 16,105 | |
| currency | | 99822 | | 3 | | USD | | 99820 | | 0,004 | |
| fee | | 99823 | | 5 | | No | | 99619 | | 0,003 | |
| has\_photo | | 99823 | | 6 | | Yes | | 56165 | | 0,003 | |
| pets\_allowed | | 39192 | | 6 | | Cats,Dogs | | 37213 | | 60,740 | |
| price\_type | | 99823 | | 6 | | Monthly | | 99816 | | 0,003 | |
| address | | 7946 | | 7774 | | 8215 S.W 72nd Avenue | | 19 | | 92,040 | |
| cityname | | 99521 | | 2984 | | Dallas | | 2862 | | 0,306 | |
| state | | 99521 | | 53 | | TX | | 11279 | | 0,306 | |
| source | | 99820 | | 25 | | RentDigs.com | | 91239 | | 0,006 | |
| price\_display | | 99820 | | 4546 | | 1350 | | 571 | | 0,006 | |
| **Feature** | **Mode** | | **Mode Freq** | | **Mode %** | | **2nd Mode** | | **2nd Mode Freq** | | **2nd Mode %** |
| category | housing/rent/apartment | | 99762 | | 99,936 | | housing/rent/commercial/retail | | 42 | | 0,042 |
| title | Apartment in great location | | 1066 | | 1,068 | | Apartment in prime location | | 1031 | | 1,033 |
| body | When searching for a pet-friendly... | | 45 | | 0,045 | | Welcome to Apartment homes... | | 37 | | 0,037 |
| amenities | Parking | | 6204 | | 6,215 | | Parking,Storage | | 2117 | | 2,121 |
| currency | USD | | 99820 | | 99,994 | | Cats,Dogs | | 1 | | 0,001 |
| fee | No | | 99619 | | 99,793 | | Yes | | 201 | | 0,201 |
| has\_photo | Yes | | 56165 | | 56,263 | | Thumbnail | | 34394 | | 34,454 |
| pets\_allowed | Cats,Dogs | | 37213 | | 37,278 | | Cats | | 1848 | | 1,851 |
| price\_type | Monthly | | 99816 | | 99,990 | | Weekly | | 3 | | 0,003 |
| address | 8215 S.W 72nd Avenue | | 19 | | 0,019 | | 2647 Eastgate Road | | 8 | | 0,008 |
| cityname | Dallas | | 2862 | | 2,867 | | Denver | | 2760 | | 2,765 |
| state | TX | | 11279 | | 11,299 | | CA | | 10341 | | 10,359 |
| source | RentDigs.com | | 91239 | | 91,398 | | RentLingo | | 6924 | | 6,936 |
| price\_display | 1350 | | 571 | | 0,572 | | 850 | | 542 | | 0,543 |

Buvo panaikinti visi įrašai, kurių valiuta nėra „USD“, kaip pvz. 3 lentelės antroje modoje valiuta „Cats,Dogs“. Analizuojant 3 lentelę, buvo kilę įtarimų dėl dublikatinių įrašų sprendžiant pagal pirmąją ir antrąją „body“ modą, tačiau perskaičius turinį, panašu, kad tai yra aprašas nuomojamų būtų komplekse, todėl įrašus nuspręsta palikti.

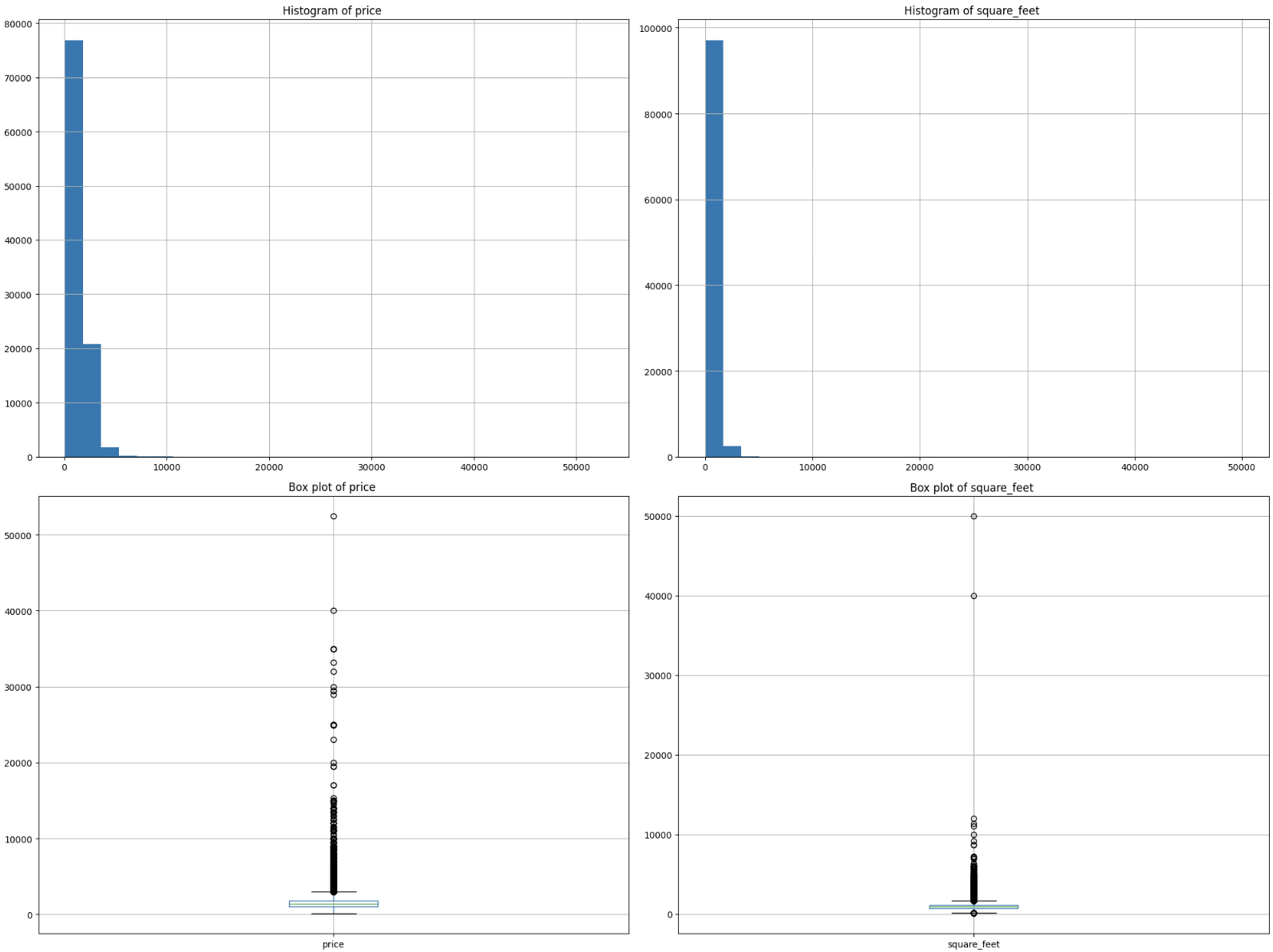
## Distribution Characteristics

Toliau buvo analizuojamas pasiskirstymo charakteristikos nusibrėžiant histogramas, dėžės ir pyrago diagramas. Žemiau pateiktos pagrindinių 4 ypatybių – vonios kambarių skaičius, miegamųjų skaičius, ploto dydis bei kaina, diagramos. Iš 1 paveikslo diagramos matyti, kad 2 didžiausi skaičiai vonios ir miegamųjų kambarių yra 1 arba 2. Vonių kambarių dažniausiai būna vienas, o miegamųjų 2. Taip pat matoma, kad yra ir išimčių, miegamųjų gali būti daugiau ir mažiau (0 ir 9), o vonios kambarių daugiau (1 ir 9). Išimtys ypatingai matomos, dėžės diagramose (t.y. juodais apskritimais pažymėti taškai).



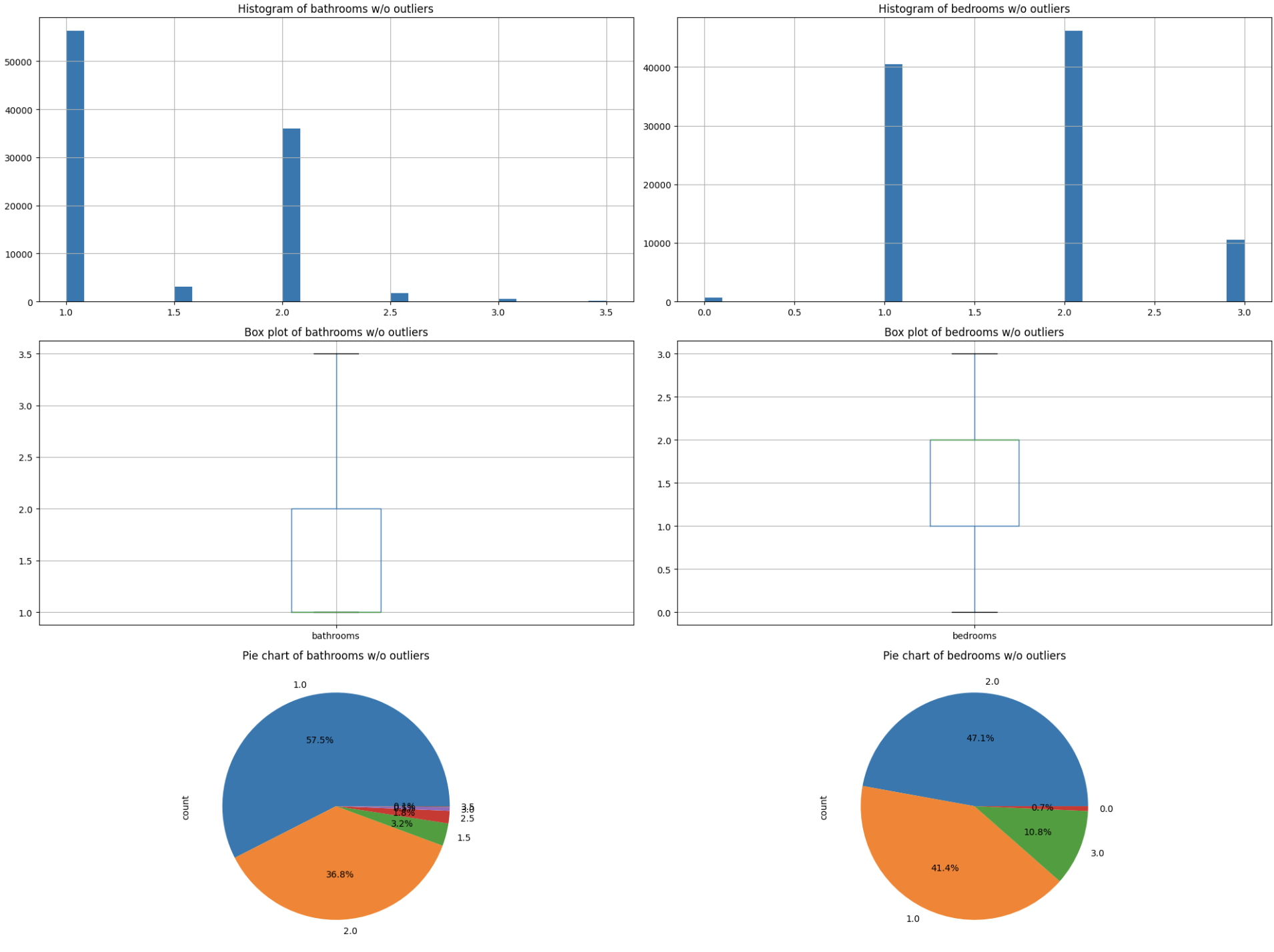
1. pav. Vonios ir miegamųjų kambarių pasiskirstymo diagramos

Iš 4 tiriamų savybių, didžiausi išsiskyrimai (angl. deviations) matomi kainose ir plote. Iš diagramų 2 paveiksle, matyti kaip plotas siekia 50000 kvadratinių pėdų, o kaina siekia virš 50000 JAV dolerių.



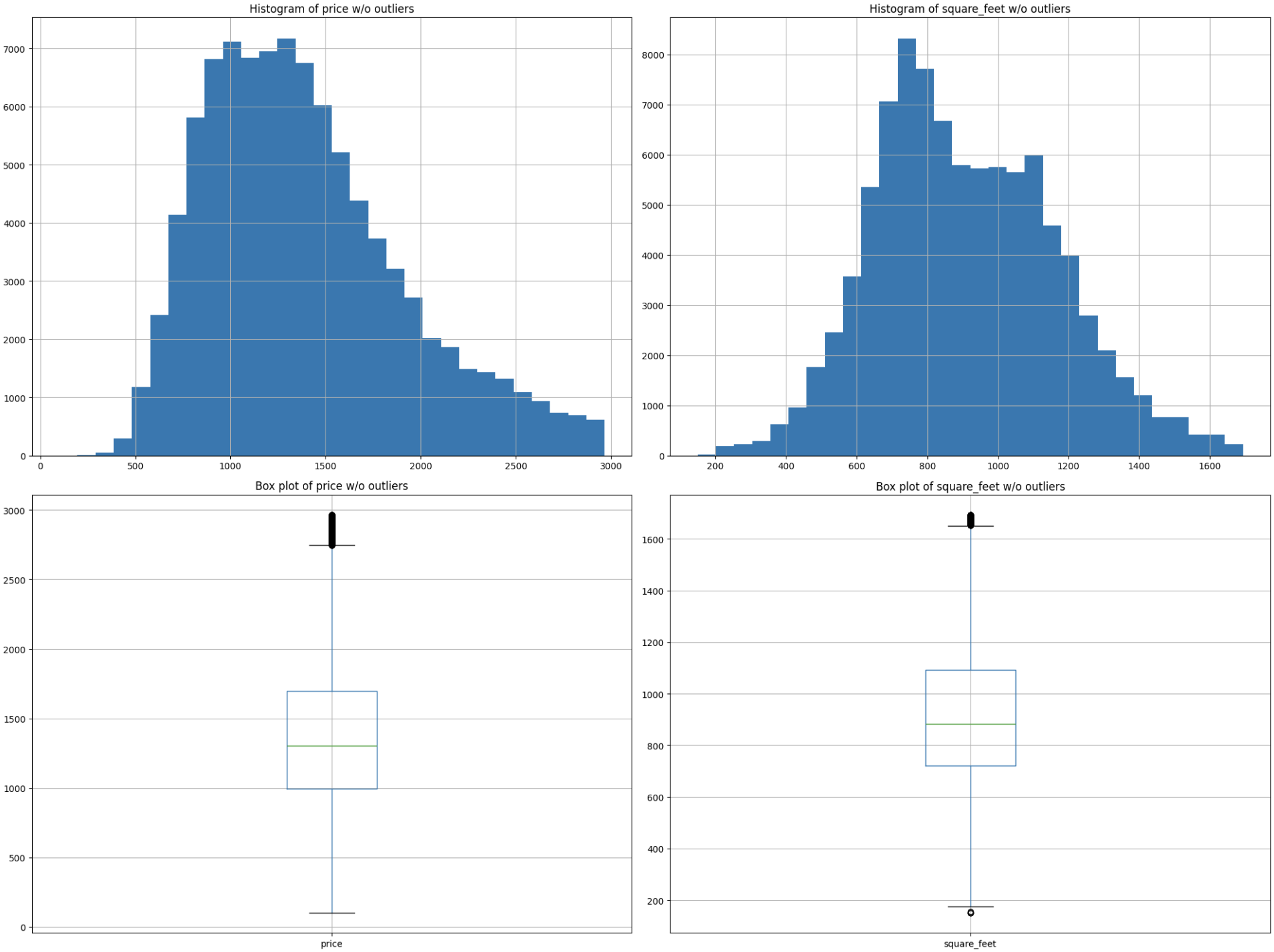
2. pav. Kainos ir ploto pasiskirstymo diagramos

Dėl didelių išsiskyrimų diagramose sunku įžvelgti pasiskirstymą tarp mažesnių verčių, kurios sudaro didžiausiu stulpelius. Dėl to, šie išsiskyrimai buvo panaikinti, pasinaudojant interkvartilinio intervalo metodu (angl. Interquantile Range Method [1]). Toliau paveiksluose 3, 4 pateikti rezultatai diagramose po šio metodo pritaikymo duomenų rinkiniui.



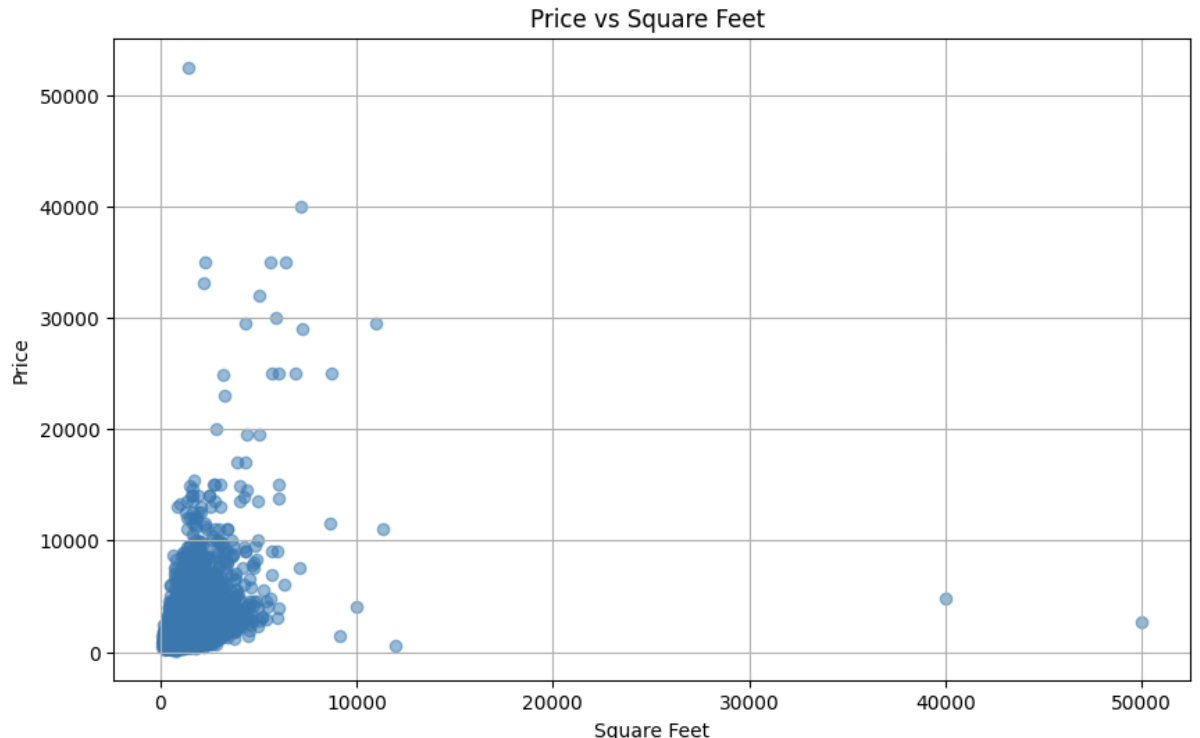
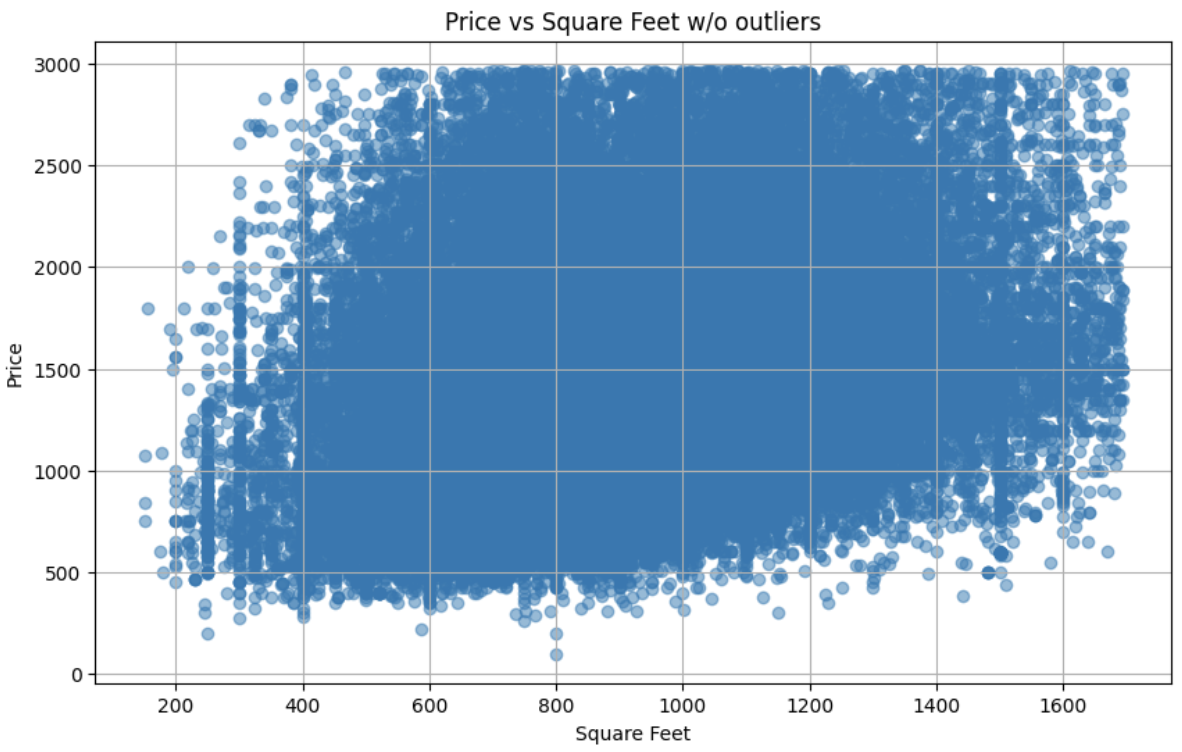
3. pav. Vonios ir miegamųjų kambarių pasiskirstymo diagramos be atsiskyrusių verčių

Po išimčių panaikinimo metodo diagramose įžvelgiame platesnį pasiskirstymą ir išsiskiriančios reikšmės dežės diagramose yra sužymėtos arčiau viena kitos (4 pav.)



4. pav. Kainos ir ploto pasiskirstymo diagramos be atsiskyrusių verčių

Papildomai pateiktos 5 pav. diagramos kainos ir ploto pasiskirstymo prieš ir po išimčių panaikinimo. Galima matyti, kaip buvo panaikintos atitolusios reikšmės.

5. pav. Kainos ir ploto pasiskirstymo diagramos prieš ir po išimčių

## Išvestinės savybės

Taip pat, galime susikurti išvestines savybes tokias kaip, ploto ir kainos, vonios ir miegamųjų kambarių santykius. Taip pat įmanomos ir dvejetainės reikšmės (angl. flags) tokios kaip, ar galima bute laikyti šunį, katę arba abu ar yra parkavimas, virtuvė. Sprendimų priėmimo modeliams išmokyti užtenka ir esamų duomenų, todėl šios reikšmės nebus pildomos.

## Duomenų paruošimas

Buvo panaikinti neteisingi įrašai, tokie, kurie neturėjo reikšmių (pvz. nenurodyta kaina), arba duomenys buvo akivaizdžiai neteisingi (pvz. klaidinga valiuta). Taip pat, panaikintos išsiskiriančios vertės. Be išimčių panaikinimo, papildomai duomenys buvo standartizuojami, kad būtų sutapatintos dydžių vertės tokios kaip kvadratinės pėdos ir miegamųjų skaičius, tam, kad skaičiuojant modelis nebūtų šališkas vienam iš didesnių parametrų. Kitiems paklaidos skaičiavimams buvo palikta originali kainos reikšmė. Prieš duomenų paruošimą buvo 99819 įrašai, po – 92431. Tolimesniam skyriuje šie duomenys prieš ir po paruošimą (palyginumui) buvo pritaikyti skirtingiems sprendimus priimantiems modeliams.

# KNN, DT, RF ir ANN implementacijos, geriausios hiperparametrų aibės ir ataskaita

Implementacijoms, taip pat ir duomenų manipuliacijai ir diagramų generavimui buvo naudojamos 1.1. poskyryje minėtos bibliotekos. Buvo naudojamos standartizuotos ypatybės – *ilguma, platuma, kaina, plotas miegamųjų ir vonios kambarių skaičius*. Taip pat, skirtingų ypatybių skaičių įtakai ištirti buvo panaikinta *ilguma ir platuma*.

## KNN ataskaita, geriausi hiperparametrai

KNN prognozė atliekama remiantis gretimomis tikslinėmis vertėmis K ir jas sudarančiais parametrais, apskaičiuojant atstumą nuo tikrosios vertės. artimiausių kaimynų regresijos modelio implenetacijai buvo panaudotas scikit learn sprendimas [2]. Šiam modeliui, kaip ir sekantiems, buvo pritaikyti skirtingi hiperparametrai K = 3, K = 10, K = 20 ir duomenų rinkinys prieš ir po paruošimą.

**4. lentelė.** Rezultatai metrikomis MAE, MAPE ir RMSE, su skirtingomis hiperparametro *K* vertėmis ir paruoštais ir neparuoštais duomenimis. Pasvirusiai paryškintas geriausias rezultatas.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Duom. Paruošti?** | **K** | **mae, $** | **mape** | **rmse, $** | **time, s** |
| Taip | ***3*** | ***184,33*** | ***0,142*** | ***280,22*** | ***0,632*** |
| 10 | 208,42 | 0,162 | 291,94 | 0,655 |
| 20 | 224,77 | 0,176 | 306,50 | 0,664 |
| Ne | 3 | 245,20 | 0,155 | 648,18 | 0,652 |
| 10 | 270,27 | 0,178 | 646,27 | 0,606 |
| 20 | 286,63 | 0,187 | 649,76 | 0,675 |
| Taip, be ilgumos ir platumos | 3 | 423,18 | 0,337 | 547,08 | 0,635 |
| 10 | 388,97 | 0,311 | 499,10 | 0,466 |
| 20 | 382,71 | 0,307 | 489,59 | 0,499 |

Iš 4 lentelėje pateiktų rezultatų su duomenimis, kurie turėjo daugiau ypatybių, matome, kad kuo mažesnis K, tuo modelio tikslumas yra didesnis, t.y. paklaidos metrikos yra mažesnės, taip pat ir laikas mažesnis, nes atliekami mažesni skaičiavimai vienai vertei prognozuoti.

Panaikinus, ilgumą ir platumą yra mažiau ypatybių, kurias modelis naudoja apskaičiuoti kainą, todėl matoma, kad tikslumas ne vien yra prastesnis už duomenis prieš ir po paruošimo, bet ir K turi būti didesnis siekiant mažesnės paklaidos, tam, kad būtų apskaičiuotas bendresnis rezultatas. Dėl mažesnio ypatybių skaičiaus, taip pat skaičiavimai užtrunka mažiau.

Nors galime, matyti iš 4 lentelės rezultatų, kad K = 3 yra optimaliausias hiperparametras. Kad įsitikinti, papildomai buvo atliktas tyrimas, pritaikant parametrų tinklo paiešką su kryžmine validacija (angl. Grid Search Cross-Validation [3]). Žemiau 5 lentelėje, pateikti kryžminės validacijos su 3 duomenų segmentais (angl. folds) rezultatai, kurių taškai buvo skaičiuojami pagal neigiamą absoliutų vidurkio paklaidą (negative mean absolute error). Tyrimas patvirtina, kad rezultatai su K = 3 turi mažiausią paklaidą, kai naudojami paruošti duomenys su visais skyriaus įvade minėtais parametrais.

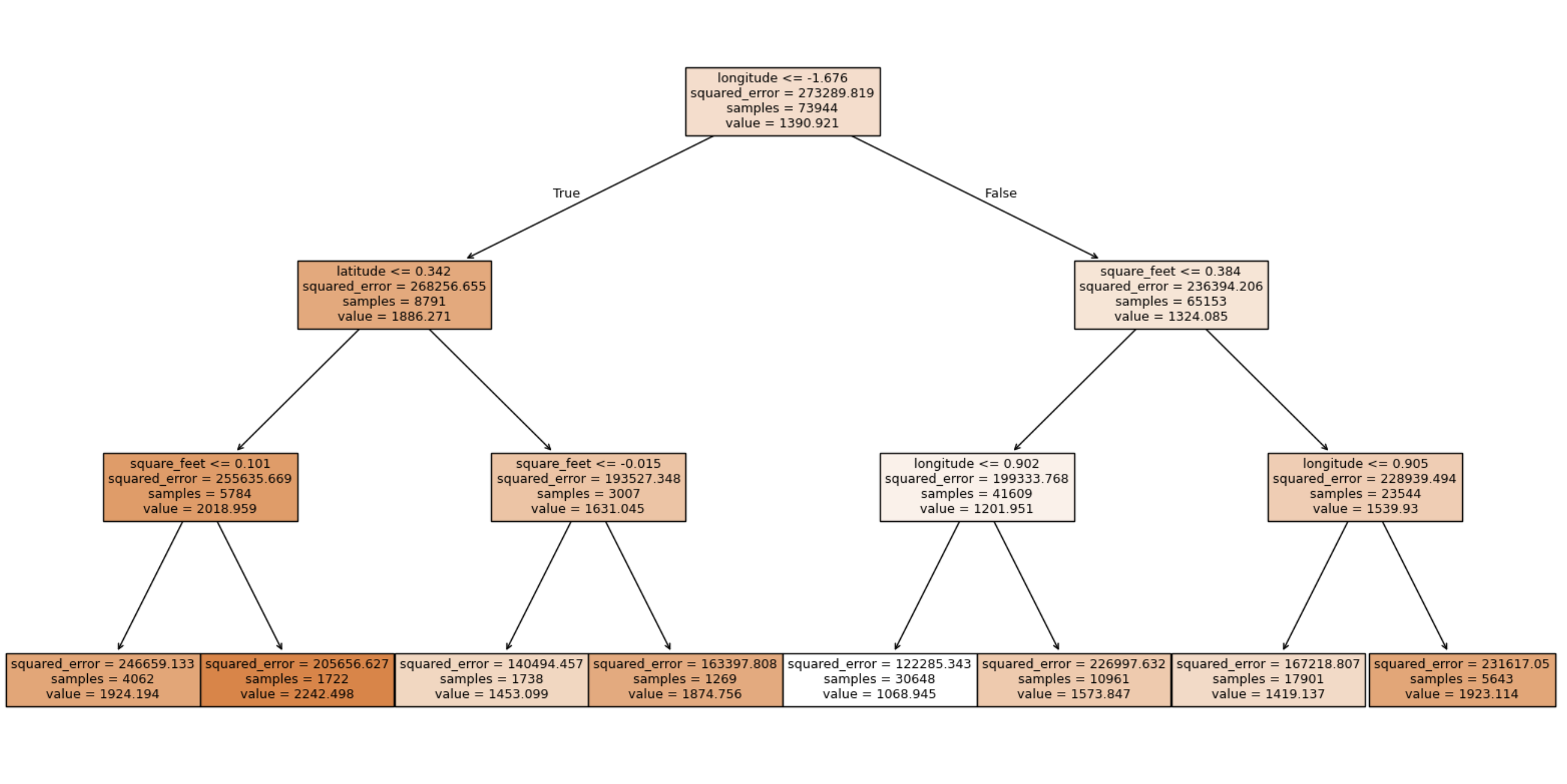
**5. lentelė.** Geriausio K paieškos kryžminės validacijos rezultatai, matoma, kad K = 3, turi mažiausia neigiamą testo taškų skaičių. Pasvirusiai paryškintas geriausias rezultatas.

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **neg\_mean\_test\_score** | **mean\_score\_time, s** | **mean\_fit\_time, s** |
| ***3*** | ***-0,070*** | ***0,420*** | ***0,025*** |
| 5 | -0,073 | 0,638 | 0,034 |
| 10 | -0,077 | 0,724 | 0,032 |
| 15 | -0,080 | 0,661 | 0,029 |
| 20 | -0,082 | 0,705 | 0,028 |

## DT ataskaita, geriausi hiperparametrai

Sprendimų medžio atšakos generuojamos pagal kriterijus (žr. 6 pav.) , kurie labiausiai padidina gaunamos informacijos kiekį ir sumažina paklaidą. Sprendimų medžiui taip pat buvo panaudota scikit implementacija [4]. Naudojami hiperparametrai buvo:

* Maksimalus gylis (max\_depth);
* Minimalus mėginio dydis lapui (min\_samples\_leaf);
* Kainos ir kompleksiškumo daugiklis (ccp\_alpha);
* Maksimalus ypatybių skaičius medyje (max\_features).



6. pav. 3 gylių sprendimų medis, sudarytas pagal daugiausiai informacijos suteikiančias ypatybės ir jų slenksčių vertes

Atlikta hiperparametrų įtakos modelio tikslumui analizė, matoma 6 lentelėje. Nuspręsta palikti maksimalių ypatybių skaičių kaip konstantą, kad būtų mažiau kintamųjų tyrime ir būtų galima daryti tikslesnes išvadas.

**6. lentelė.** Sprendimų medžio paklaidos ir greičio rezultatai su trijų skirtingų aibių hiperparametrais *max\_depth, min\_samples\_leaf* ir *ccp\_alpha*. Pasvirusiai paryškintas geriausias rezultatas.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Duom. Paruošti?** | **max\_**  **features** | **max\_**  **depth** | **min\_**  **samples\_**  **leaf** | **ccp\_**  **alpha** | **mae, $** | **mape** | **rmse, $** | **time, s** |
| Taip | 3 | 3 | 100 | 0,01 | 319,55 | 0,26 | 410,01 | 0,027 |
| 5 | 200 | 0,005 | 288,53 | 0,23 | 375,16 | 0,057 |
| ***7*** | ***400*** | ***0,001*** | ***269,83*** | ***0,21*** | ***355,60*** | ***0,089*** |
| Ne | 3 | 100 | 0,01 | 419,50 | 0,29 | 785,89 | 0,048 |
| 5 | 200 | 0,005 | 380,76 | 0,26 | 735,33 | 0,061 |
| 7 | 400 | 0,001 | 351,40 | 0,24 | 720,82 | 0,079 |
| Taip, be ilgumos ir platumos | 3 | 100 | 0,01 | 395,50 | 0,32 | 497,87 | 0,034 |
| 5 | 200 | 0,005 | 391,02 | 0,32 | 491,77 | 0,041 |
| 7 | 400 | 0,001 | 390,80 | 0,32 | 491,60 | 0,076 |

6 lentelėje matomas akivaizdus skirtumas RMSE metrikoje su neparuoštais duomenimis, kurie turi daug kraštutinumų, šie padidina modelio paklaidą. RMSE išryškina spėjimus toli nuo tiesos. Taip galima pastebėti, kad dėsningai didesnis maksimalus medžio gylis, didesnis mėginių dydis paskutiniems mazgams (lapams) ir mažesnis kainos-kompleksiškumo karpymo daugiklis veda prie tikslesnių rezultatų, tačiau modelio generavimas ir prognozavimas užtrunka ilgiau. Paskutinėse 3 eilutėse, galime matyti, kad esant mažiau ypatybių modelio tikslumas kenčia.

Tolimesniame žingsnyje buvo atliekama hiperparametrų optimizacija tuo pačiu būdu pritaikant kryžminę validaciją hiperparametrų tinklui kaip ir 3.1. KNN poskyryje.

**7. lentelė.** Sprendimų medžio paklaidos ir greičio rezultatai su trijų skirtingų aibių hiperparametrais *max\_depth, min\_samples\_leaf* ir *ccp\_alpha*. Su ccp\_alpha = 0,01, dėl mažos įtakos ir max\_features = 3. Geriausias rezultatas pasviręs ir paryškintas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **max\_depth** | **min\_samples\_leaf** | **mean\_test\_score** | **mean\_score\_time** | **mean\_fit\_time** |
| 3 | 100 | -320,029 | 0,002 | 0,034 |
| 200 | -320,029 | 0,002 | 0,030 |
| 400 | -320,029 | 0,002 | 0,031 |
| 5 | 100 | -290,028 | 0,002 | 0,042 |
| 200 | -290,050 | 0,002 | 0,062 |
| 400 | -289,729 | 0,002 | 0,041 |
| ***7*** | ***100*** | ***-263,122*** | ***0,003*** | ***0,059*** |
| 200 | -265,809 | 0,003 | 0,051 |
| 400 | -272,552 | 0,002 | 0,063 |

Tiksliausias sprendimas gaunamas taikant didesnį gylį ir mažesnį imties dydį.

## RF ataskaita, geriausi hiperparametrai

Atsitiktinio miško algoritmas, naudoja sprendimų medžio idėja sukurdamas daugiau nei vieną sprendimų medį atsitiktine tvarka, išrinkdamas atsitiktines ypatybes ir duomenų imtis [5]. Vėliau apskaičiuojamas kiekvieno medžio išvesties vidurkis. Nagrinėjami šie hiperparametrai:

* Medžių skaičius (n\_estimators);
* Minimalus mėginio dydis lapui (min\_samples\_leaf);

**8. lentelė.** Atstitiktinio miško paklaidos ir greičio rezultatai, naudojant medžių skaičiaus ir minimalaus mėginio dydžio lapui hiperparametrus. Pasvirusiai paryškintas geriausias rezultatas.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Duom. Paruošti?** | **n\_**  **estimators** | **min\_**  **samples\_**  **leaf** | **mae** | **mape** | **rmse** | **time, s** |
| ***Taip*** | ***25*** | ***100*** | ***213,02*** | ***0,17*** | ***286,95*** | ***1,885*** |
| 50 | 200 | 229,04 | 0,18 | 306,62 | 2,918 |
| 100 | 500 | 250,90 | 0,20 | 332,18 | 5,288 |
| Ne | 25 | 100 | 271,03 | 0,18 | 636,04 | 2,027 |
| 50 | 200 | 292,40 | 0,20 | 664,09 | 3,049 |
| 100 | 500 | 319,75 | 0,22 | 696,40 | 4,777 |
| Taip, be ilgumos platumos | 25 | 100 | 385,40 | 0,31 | 486,67 | 0,934 |
| 50 | 200 | 387,64 | 0,31 | 488,33 | 1,314 |
| 100 | 500 | 390,49 | 0,32 | 491,06 | 2,041 |

Išmatavus paklaidas, 8 lentelėje matoma, kad tiksliausius spėjimus iš visų modelių galime gauti su atstiktinio miško algortimu. Taip pat, matome, kad mažesnis skaičius medžių ne tik sumažina skaičiavimo laiką, bet ir yra tikslesnis, matomai dėl to, kad kitais atvejais modelis „persimoko“. Duomenų paruošimas, kaip ir praeituose matavimuose, turi didelės įtakos modelio tikslumui, ypač RMSE metrikai, kuris išryškina labiausiai nutolusias prognozes. Kitų metrikų prasčiausi rezultatai buvo pasiekti, kada buvo naudojama duomenų rinkinys su mažiau ypatybių.

**9. lentelė.** Atsitiktinio miško hiperparametrų optimizavimas, didžiausias neigiamas taškų skaičius gautas su min\_samples\_leaf = 100 ir n\_estimators = 500.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **min\_samples\_leaf** | **n\_**  **estimators** | **mean\_test\_score** | **mean\_score\_time** | **mean\_fit\_time** |
| 100 | 100 | -222,478 | 0,043 | 0,909 |
| 200 | -221,859 | 0,079 | 1,716 |
| ***500*** | ***-221,628*** | ***0,157*** | ***3,387*** |
| 200 | 100 | -238,491 | 0,034 | 0,763 |
| 200 | -237,878 | 0,068 | 1,504 |
| 500 | -237,457 | 0,132 | 2,993 |
| 500 | 100 | -262,645 | 0,027 | 0,607 |
| 200 | -262,044 | 0,052 | 1,217 |
| 500 | -261,646 | 0,108 | 2,405 |

Geriausias rezultatas buvo pasiektas su didžiausiu skaičium medžių, tačiau priimant laiko kompromisą, kuris buvo ilgiausias iš visų kitų bandymų.

## ANN ataskaita, geriausi hiperparametrai

ANN implementacijai buvo pasitelktas daugiasluoksnis perceptronas (angl. Multi-Layer Perceptron [6]). Šis sprendimas susidaro iš įvesties, paslėptojo ir išvesties sluoksnių. Paslėptasis sluoksnis gali susidaryti iš skirtingo skaičiaus perceptronų sluoksnių bei pačių perceptronų. Kiekvienas iš jų priima kiekvieną įvestį ir atlieka skaičiavimus kartu su svoriais bei šališkumo vertėmis. Skaičiavimai vėliau yra pritaikomi aktyvacijos funkcijai prieš perduodant į sekančius sluoksnius. Išvesties ir tiesos skirtumas yra apskaičiuojamas pritaikant praradimo funkciją, tokią kaip MSE ir, pasitelkiant galiniu propagavimu, svoriai yra optimizuojami gradiento nusileidimo algoritmu. Šis procesas yra kartojamas, kol pasiekiama konvergacija arba iteracijų limitas. Iteracijų limitas nebuvo keičiamas ir buvo paliktas ties 200 riba.

Analizėje naudojami hiperparametrai buvo:

* Aktyvacijos funkcija (activation);
* Paslėpti sluoksniai ir jų dydžiai (hidden\_layer\_sizes);
* Mokymosi daugiklis, kuris buvo paliktas kaip konstanta (learning\_rate\_init).

**10. lentelė.** Daugiasluoksnio perceptrono tyrimo rezultatai. hidden\_layer\_sizes pirmas skaičius reiškia pirmo sluoksnio perceptronų skaičių, antras – antro. Pasvirusiai paryškintas geriausias rezultatas.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Duom. Paruošti?** | **activation** | **hidden\_**  **layer\_sizes** | **learning\_**  **rate\_init** | **mae** | **mape** | **rmse** | **time, s** |
| ***Taip*** | ***relu*** | ***(20, 10)*** | ***0.01*** | ***254,41*** | ***0,20*** | ***339,54*** | ***35,271*** |
| (30, 15) | 0.001 | 259,03 | 0,20 | 346,58 | 42,838 |
| tanh | (10, 5) | 0.01 | 268,57 | 0,21 | 355,27 | 41,468 |
| Ne | relu | (20, 10) | 0.01 | 358,25 | 0,25 | 700,30 | 50,251 |
| (30, 15) | 0.001 | 366,62 | 0,25 | 717,31 | 48,011 |
| tanh | (10, 5) | 0.01 | 348,69 | 0,23 | 731,76 | 44,270 |
| Taip, be ilgumos platumos | relu | (20, 10) | 0.01 | 388,88 | 0,31 | 492,68 | 5,530 |
| (30, 15) | 0.001 | 391,37 | 0,32 | 491,11 | 53,289 |
| tanh | (10, 5) | 0.01 | 417,68 | 0,35 | 524,59 | 19,404 |

Matoma kaip didesnis mokymosi daugiklis padeda pasiekti konvergacijos reikšmę greičiau. Mokymosi daugiklio įtaką galime ypatingai pastebėti priešpaskutinėje eilutėje. 9 lentelėje nustatyta 2 sluoksniai su 20 ir 10 perceptronų kaip optimaliausia hiperparametrų aibė. Papildomai, matoma, kad relu yra tinkamesnė aktyvacijos funkcija, tai gali būti dėl to, kad tanh išvesties reikšmės apsiriboja (-1;1) intervalu, todėl gradiento reikšmės naudojamos svoriams keisti yra mažos. Taip pat pasikartoja jau anksčiau matyti rezultatai su neparuoštais duomenimis – tikslumas mažesnis.

Žemiau 10 lentelėje, kaip ir su kitais modeliais, buvo atlikta atranka su kryžmine validacija geriausių hiperparametrų kombinacijų paruoštiems duomenims.

**11. lentelė.** Daugiasluoksnio perceptrono geriausių hiperparametrų atranka. Tiksliausias sprendimas pasvirusiai paryškintas buvo su 20 ir 10 perceptronų sluoksniais bei smulkesniu 0,001 mokymosi daugikliu, tačiau tai buvo lėtesnis rezultatas nei panašaus tikslumo su 0,01 daugikliu.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **activation** | **hidden\_ layer\_sizes** | **learning\_ rate\_init** | **mean\_test\_ score** | **mean\_score\_ time** | **mean\_fit\_time** |
| ***relu*** | ***(20, 10)*** | 0,01 | -393,765 | 0,007 | 5,074 |
| ***0,001*** | ***-390,885*** | ***0,006*** | ***22,327*** |
| (30, 15) | 0,01 | -396,579 | 0,009 | 10,123 |
| 0,001 | -391,758 | 0,015 | 32,667 |
| (10, 5) | 0,01 | -392,764 | 0,005 | 8,457 |
| 0,001 | -392,352 | 0,006 | 23,562 |
| tanh | (20, 10) | 0,01 | -408,265 | 0,039 | 20,884 |
| 0,001 | -846,966 | 0,037 | 30,295 |
| (30, 15) | 0,01 | -407,815 | 0,054 | 21,621 |
| 0,001 | -629,256 | 0,056 | 41,331 |
| (10, 5) | 0,01 | -399,966 | 0,020 | 20,545 |
| 0,001 | -1090,772 | 0,020 | 23,297 |

# Apibendrinimas ir išvados

Buvo atlikta duomenų analizė ir paruošimas. Rasta duomenų, kuriems trūko būtinų įrašų, tokių kaip kainos suma ar valiuta. Šie įrašai buvo panaikinti. Taip pat buvo panaikintos išimtinės, atsiskyrusios vertės (angl. outliers) pritaikant interkvartilinio intervalo metodą. Taip pat duomenys buvo standartizuojami, tam, kad skirtingi verčių dydžiai turėtų tą pačią įtaką prognozuojamam rezultatui. Prieš ir po paruošimą duomenų įrašų buvo 99819 ir 92431. Vėliau sprendimų priėmimo modelių implementacijose buvo atlikta duomenų apdorojimo bei kitų hiperparametrų įtakos analizė. Taip pat nustatyti optimaliausi hiperparametrai pagal parametrų tinklo paieškos technika su kryžmine validacija, kurioje duomenys buvo skaidomi į 3 lygias dalis. Pagrindinės ypatybės buvo naudojamos ilguma, platuma, kvadratinės pėdos, miegamųjų ir vonios kambarių skaičius.

Išvadose atrasta, kad visuose modeliuose paruošti duomenys pagerina tiek paklaida pagal RMSE kriterijus, kurie ypatingai pabrėžia atsiskyrusių duomenų įtaką, tiek patį skaičiavimo greitį. Jeigu duomenys nėra standartizuojami, tada modelis bus labiau šališkas tai ypatybei, kurios skaitinė vertė yra didesnė. Taip pat, pastebėta, kad palankiau yra turėti daugiau ypatybių, nes didesnis jų skaičius padidina spėjimo tikslumą, tai buvo matoma iš rezultatų su ir be platumos, ilgumos.

KNN nustatyta, kad tiksliam rezultatui geriausia turėti mažą K reikšmę, skaičiavimai su mažu K irgi yra greitesni. Analizuojant DT su kryžmine validacija, nustatyta, kad turėti didesnį gylį ir mažesnį minimalų mėginių skaičių galutiniams lapams yra palankiausia tiksliausiam rezultatui, tačiau tuo pačiu tai ir užtrunka ilgiau. Atsitiktinio miško rezultatai buvo geriausi su mažesniu skaičiumi medžių ir mėginių minimumu. Tuo pačiu šis modelis buvo geriausias iš visų, todėl šiam regresijos uždaviniui spręsti su būtent taip apdorotais duomenimis reikėtų rinktis šį modelį. Taip pat buvo atlikta dirbtinio neuroninio tinklo analizė pasinaudojant daugiasluoksnio perceptrono implementaciją. Nors rezultatai buvo pakankamai tikslūs ir panašūs į kitų modelių, tačiau dėl didelio skaičiavimų skaičiaus modelio mokymui daugiasluoksnio perceptrono sprendimas buvo lėčiausias.

Šaltiniai

[1] <https://procogia.com/interquartile-range-method-for-reliable-data-analysis/>

[2] <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>

[3] <https://scikit-learn.org/stable/modules/cross_validation.html>

[4] <https://scikit-learn.org/stable/modules/tree.html#regression>

[5] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

[6] <https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html>