

**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**Studijų modulio**

**DIDŽIŲJŲ DUOMENŲ RINKINIŲ TYRYBOS METODAI**

Kursinis darbas

Atliko: MGDVDA-9 grup. stud.

Edgaras Dvariškis

Martynas Jonaitis

Tikrino: lekt. Darius Aliulis

**KAUNAS, 2019**

**Turinys**

[1. Duomenų paruošimas 3](#_Toc27277557)

[1.1. Duomenų apžvalga 3](#_Toc27277558)

[1.2. Duomenų kokybės gerinimas 5](#_Toc27277559)

[1.3. DUOMENŲ AGREGAVIMAS IR APJUNGIMAS 6](#_Toc27277560)

[2. DUOMENŲ KLASTERIZAVIMAS 6](#_Toc27277561)

# Duomenų paruošimas

## Duomenų apžvalga

Projekte buvo naudojami duomenys apie mobilaus ryšio operatoriaus vartotojus. Duomenų faile yra 65 kintamieji, kurie kiekvienas turi 13844 įrašus. Kiekvieną unikalų vartotoją apibūdina trys kintamieji: *<user\_account\_id>*, *<year>* ir *<month>.* Šie kintamieji nurodo kurio periodo ir kurio vartotojo duomenys pateikti. Likę 62 kintamieji apibūdina periodo rodmenis. Juos galima suskirstyti taip:

|  |  |  |
| --- | --- | --- |
| Kintamieji | Apibūdinimas | Skaitinė reikšmė |
| calls\_outgoing\_count | Išeinančių skambučių kiekis | interger |
| calls\_outgoing\_duration | Išeinančių skambučių trukmė | numeric |
| calls\_outgoing\_duration\_max | Ilgiausias skambutis per periodą | numeric |
| calls\_outgoing\_inactive\_days | Dienų, be išeinančių skambučių, kiekis | interger |
| calls\_outgoing\_spendings | Išlaidos išeinantiems skambučiams | numeric |
| calls\_outgoing\_spendings\_max | Brangiausias skambutis per periodą | numeric |
| calls\_outgoing\_to\_abroad\_count | Skambučių į kitas šalis kiekis | interger |
| calls\_outgoing\_to\_abroad\_duration | Trukmė skambučių į uždienį | numeric |
| calls\_outgoing\_to\_abroad\_inactive\_days | Dienų, be išeinančių skambučių į kitas šalis, kiekis | interger |
| calls\_outgoing\_to\_abroad\_spendings | Išlaidos skambučiams į užsienį | numeric |
| calls\_outgoing\_to\_offnet\_count | Skambučių įkitus tinklus, kiekis | interger |
| calls\_outgoing\_to\_offnet\_duration | Išeinančių skambučių į kitus tinklus trukmė | interger |
| calls\_outgoing\_to\_offnet\_inactive\_days | Dienų, be išeinančių skambučių į kitus tinklus, kiekis | interger |
| calls\_outgoing\_to\_offnet\_spendings | Kaina, išleista skambučiams į kitus tinklus | interger |
| calls\_outgoing\_to\_onnet\_count | Skambučių, tinklo viduje, kiekis | interger |
| calls\_outgoing\_to\_onnet\_duration | Trukmė skambučių tinklo viduje | numeric |
| calls\_outgoing\_to\_onnet\_inactive\_days | Dienų, be išeinančių skambučių tinklo viduje, kiekis | interger |
| calls\_outgoing\_to\_onnet\_spendings | Išlaidos skambučiams tinklo viduje | numeric |
| gprs\_inactive\_days | Dienų be duomenų naudojimo kiekis | interger |
| gprs\_session\_count | Prisijungimų skaičius | interger |
| gprs\_spendings | Išlaidos už duomenų naudojimą | numeric |
| gprs\_usage | Išnaudotų kb kiekis | numeric |
| last\_100\_calls\_outgoing\_duration | Paskutinių 100 dienų išeinančių skambučių trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_abroad\_duration | Paskutinių 100 dienų išeinančių skambučių š kitas šalis trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_offnet\_duration | Paskutinių 100 dienų išeinančių skambučių į kitus tinklus trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_onnet\_duration | Paskutinių 100 dienų išeinančių skambučių tiklo viduje trukmė | numeric |
| last\_100\_gprs\_usage | Paskutinių 100 dienų naudotas kb kiekis | numeric |
| last\_100\_reloads\_count | Paskutinių 100 dienų papildymų skaičius | interger |
| last\_100\_reloads\_sum | Paskutinių 100 dienų papildymų suma | numeric |
| last\_100\_sms\_outgoing\_count | Paskutinių 100 dienų SMS žinučių kiekis | interger |
| last\_100\_sms\_outgoing\_to\_abroad\_count | Paskutinių 100 dienų SMS žinučių kiekis į kitas šalis | interger |
| last\_100\_sms\_outgoing\_to\_offnet\_count | Paskutinių 100 dienų SMS žinučių kiekis į kitus tinklus | interger |
| last\_100\_sms\_outgoing\_to\_onnet\_count | Paskutinių 100 dienų SMS žinučių kiekis tinklo viduje | interger |
| month | Mėnesis, kada pradėjo klientas naudotis telefono kompanijos paslaugomis | interger |
| reloads\_count | Papildymų kiekis | interger |
| reloads\_inactive\_days | Dienų, be papildymų, skaičius | interger |
| reloads\_sum | Papildymų bendra suma | numeric |
| sms\_incoming\_count | Įeinančių SMS žinučių skaičius | interger |
| sms\_incoming\_from\_abroad\_count | Įeinančių SMS žinučių iš kitų tinklų skaičius | interger |
| sms\_incoming\_from\_abroad\_spendings | Išlaidos įeinančioms SMS žintumės iš kitų šalių | numeric |
| sms\_incoming\_spendings | Išlaidos įeinančioms SMS žinutėms | numeric |
| sms\_outgoing\_count | Išeinančių SMS žinučių kiekis | interger |
| sms\_outgoing\_inactive\_days | Dienų, be išeinančių SMS žinučių, kiekis | interger |
| sms\_outgoing\_spendings | Išlaidos išeinančios SMS žinutėms | numeric |
| sms\_outgoing\_spendings\_max | Brangiausia SMS žinutė | numeric |
| sms\_outgoing\_to\_abroad\_count | Išeinančių SMS žinučių į kitas šalis skaičius | interger |
| sms\_outgoing\_to\_abroad\_inactive\_days | Dienų, be išeinančių SMS žinučių į kitas šalis, skaičius | interger |
| sms\_outgoing\_to\_abroad\_spendings | Išlaidos išeinančioms SMS žinutėms į kitas šalis | numeric |
| sms\_outgoing\_to\_offnet\_count | Išeinančių SMS žinučių į kitus tinklus kiekis | interger |
| sms\_outgoing\_to\_offnet\_inactive\_days | Dienų, be išeinančių SMS žinučių į kitus tinklus, skaičius | interger |
| sms\_outgoing\_to\_offnet\_spendings | Išlaidos išeinančioms SMS žinutėms į kitus tinklus | numeric |
| sms\_outgoing\_to\_onnet\_count | Išeinančių SMS žinučių tinklo viduje kiekis | interger |
| sms\_outgoing\_to\_onnet\_inactive\_days | Dienų, be išeinančių SMS žinučių tinklo viduje, kiekis | interger |
| sms\_outgoing\_to\_onnet\_spendings | Išlaidos išeinančioms SMS žinutėms tinklo viduje | numeric |
| user\_account\_balance\_last | Kliento sąskaitos likutis periodo gale | numeric |
| user\_account\_id | Unikalus kliento indentifikavimo numeris | interger |
| user\_does\_reload | Klientas padarė bent vieną papildymą | interger |
| user\_has\_outgoing\_calls | Klientas padarė bent vieną skambutį | interger |
| user\_has\_outgoing\_sms | Klientas išsiuntė bent vieną SMS žinutę | interger |
| user\_intake | Naujas klientas | interger |
| user\_lifetime | Kliento buvimo laikas dienomis | interger |
| user\_no\_outgoing\_activity\_in\_days | Dienų, kai vartotojas neatliko jokių operacijų, kiekis | interger |
| user\_spendings | Išlaidos per periodą | numeric |
| user\_use\_gprs | Klientas panaudojo duomenis bent vieną kartą | interger |
| year | Metai, kada pradėjo klientas naudotis telefono kompanijos paslaugomis | interger |

28 *numeric* (realieji skaičiai) tipo kintamieji

34 *interger* (sveikieji skaičiai) tipo kintamieji

## **Duomenų kokybės gerinimas**

Prieš pradėdami duomenų paruošimo modeliavimui žingnius (istornių paslaugų vartojimo duomenų agregravimą pagal vartotojo ID) ir modeliavimą, turime atlikti žvalgomają analizę ir duomenų kokybės gerinimą. Duomenų kokybės gerinimas buvo vykdomas tiukrinant kiekvieną kintamąjį ir ieškant nekorektiškų reikšmių. Randant tokias rekšmes jas šaliname iš imties ir tolimesnį darbą darome su likusia imtimi. Tai darysime naudodmai python funkciją „Describe“, kuri padės matyti: minimumo ir maksimumo taškus, kvantilius, standartinį nuokrypį ir vidurkį ; naudosime histogramas, kuriuose matysis, kaip pasiskirstę duomenys. Toliau aprašysime tuos kintamuosius, kuriuos reikėjo sutvarkyti:

***user\_lifetime*** - šis kintamasis apibūdina kiek dienų klientas jau yra vartotojas. Nusibrėžus šio kintamojo histogramą matome, jog dalis reikšmių yra didesnės negu 15000. Pavertus šitas dienas į laikotarpį gauname, kad šie klientai naudojosi mobilaus ryšio kompanijos paslaugomis daugiau negu 40 metų. Pirmas komercinis mobilus ryšys rinkai buvo pristatytas 1979 metais, o mūsų duomenys yra 2013 metų. Galima teiti, kad šie įrašai yra klaidingi ir tada juos tenka pašalinti.

***user\_no\_outgoing\_activity\_in\_days*** - šis kintamasis nusako kiek dienų klientas yra neaktyvus. Pagal tokią pačią logiką, kurią naudojome kintamajam *<user\_lifetime>*. Patikrinus ar yra įrašų, kuriuose neaktyvumas yra didesnis negu vartotojo buvimas klientu, pašaliname iki 2 ir nuo 98 ekretemalias reikšmes.

***user\_account\_balance\_last*** - kliento sąskaitos likutis periodo gale. Kintamasis pasižymintis vidurkiu apie 10 ir standartiniu nuokrypiu 18 turi gan dideles maksimalias (1791) reikšmes lyginant su 95 kvantiliu(35,6). Šiam kintamajam taip pat panaikiname ekstremalias reikšmes.

***user\_spendings*** - Kintamasis nurodo vartotojo per periodą (mėnesį) pinigų išleidimą (valiuta nežinoma). Minimalios reikšmės rodo, kad yra neigiamų įrašų. Šie įrašai pašalinami kaip nekorektiški, nes išleidimas negali būti neigiamas. Taip pat pašalinamos nuo 98 procentilio reikšmes.

***reloads\_sum -*** kintamasis nurodo už kokią sumą per periodą (mėnesį) buvo nupirkta papildymų. Buvo rasta neigiamų reikšmių, kurios buvo pašalintos.

***calls\_outgoing\_duration*** - kintamasis parodo kokia per periodą (mėnesį) buvo bendra skambučių trukmė. Kintamasis pasižymintis vidurkiu apie 33 ir standartiniu nuokrypiu 63 turi gan dideles maksimalias (2910) reikšmes lyginant su 95 kvantiliu(139). Šiam kintamajam taip pat panaikiname ekstremalias reikšmes.

***calls\_outgoing\_inactive\_days -*** kintamasis nurodo, kiek dienų klientas nepadarė nei vieno skambučio. Norint pasiekti geresnį modelį pašaliname ektremalias reikšmes šiam kintamąjam.

***calls\_outgoing\_to\_onnet\_inactive\_days -*** buvo pastebėta, jog šio kintamojo reikšmės yra lygios kintamojo *< calls\_outgoing\_inactive\_days>* reikšmėms.

***calls\_outgoing\_to\_offnet\_inactive\_days*** - buvo pastebėta, jog šio kintamojo reikšmės yra lygios kintamojo *< calls\_outgoing\_inactive\_days>* reikšmėms.

***calls\_outgoing\_to\_abroad\_inactive\_days*** - buvo pastebėta, jog šio kintamojo reikšmės yra lygios kintamojo *< calls\_outgoing\_inactive\_days>* reikšmėms.

***sms\_outgoing\_spendings*** - šis kintamasis nurodo kiek buvo išleista per periodą (mėnesį) SMS žinutėms. Buvo rasta neigiamų reikšmių. Jas pašaliname.

***sms\_outgoing\_inactive\_days*** - kintamasis nurodo, kiek dienų klientas neišsiuntė nei vienos SMS žinutės. Pašaliname ekstremalias reikšmes šiam kintamajam.

***sms\_outgoing\_to\_onnet\_inactive\_days*** - kintamojo reikšmės visiškai sutampa su *< sms\_outgoing\_inactive\_days>* reikšmėmis.

***sms\_outgoing\_to\_offnet\_spendings*** - kintamasis parodo, kiek buvo išleista pinigų (nežinome kokia valiuta) SMS žinučių siuntimui į kitus tinklus. Buvo pašalintos neigiamos reikšmės.

***sms\_outgoing\_to\_offnet\_inactive\_days*** - kintamojo reikšmės visiškai sutampa su *< sms\_outgoing\_inactive\_days>* reikšmėmis.

***sms\_outgoing\_to\_abroad\_inactive\_days*** - kintamojo reikšmės visiškai sutampa su *< sms\_outgoing\_inactive\_days>* reikšmėmis.

***sms\_incoming\_spendings*** - kintamasis parodo kiek buvo išleista SMS žinutėms, kurios buvo gautos. neigiamos reikšmės buvo pašalintos.

***gprs\_usage -*** kintamasis nurodo kiek buvo išnaudota mobiliųjų duomenų per periodą (mėnesį). Buvo rasta nekorektiškų duomenų imties uodegose. Kad modelis būtų tikslesnis, buvo pašalintos ektremalios reikšmės.

***gprs\_spendings -*** kintamasis nurodo kiek buvo išleista mobiliesiems duomenims per periodą (mėnesį). Buvo pašalinti įrašai su neigiamomis reikšmėmis.

***gprs\_inactive\_days*** - kintamasis nurodo, kiek dienų klientas nesinaudojo mobiliaisiais duomenimis. Pašaliname ekstremalias reikšmes, kad gautume korektiškesnį modelį.

***last\_100\_reloads\_sum*** - kintamasis parodo kokia buvo paskutinių 10 papildymų bendra suma. Rasta ir pašalintos neigiamos reikšmės.

## Duomenų agregavimas ir sujungimas

Atlikus duomenų kokybės gerinimo žingsnį, paslaugų vartojimo duomenys suagreguojami pagal vartotojo ID, kad kiekvienam vartotojui turėtume po vieną stebėjimą. Vidutiniškai vienam vartotojui tenka po 2.1 įrašo (104009 įrašai 49693 vartotojams).

Grupavimas atliekamas tolydiems kintamiesiems skaičiuojant vidutinę reikšmę o diskretiems randant didžiausią reikšmę. Vartotojai su agreguotais įrašais sujungiami per vartotojo ID stulpelį. Pastebima, kad po šių veiksmų lieka 38806 eilutės. Tai indikuoja, kad duomenų tvarkymo etape, buvo rasta vartotojų, kurių visi įrašų kokybė buvo įvertinta kaip prasta ir jie buvo panaikinti. Sujungtos eilutės saugomos *Parquet* formatu.

Programos kodas naudotas duomenų agregavimui pateikiamas Programos kodas 1, o apjungimui Programos kodas 2.

# Duomenų klasterizavimas

## Kintamųjų atranka

Nenaudosime trijų kintamųjų, kurie apibūdina įrašą - *< user\_account\_id>, <year>* ir *<month>;*

Nenaudosime 5 kintamųjų, kurie yra binariniai: *< user\_intake>, < user\_has\_outgoing\_calls>, < user\_has\_outgoing\_sms>, < user\_use\_gprs>* ir *< user\_does\_reload>.*

Kadangi klasterius sudarinėsime remdamiesi periodo duomenimis, nenaudosime ir 11 kintamųjų, kurie nurodo kliento paskutinių 100 dienų vartojimų skaičiavimus.

Taip pat, darome prielaidą, jog vartotojams yra svarbu paslaugos, kuriomis jie gali naudotis, o gaunamų SMS žinučių rodikliai į tai neįeina, todėl 4 kintamųjų, kurie tai ir apibūdina – taip pat nenaudosime.

Analizuojant kintamuosius buvo pastebėta, jog kintamieji rodantys neaktyvumą dubliuojasi, todėl paliksime tik po vieną unikalų kintamąjį (ir 6 nenaudosime kaip dublikatų).

Likusiems 36 suskaičiavome Pearsono koreliacijos koeficientus ir pašalinome kintamuosius iš tolimesnės analizės. Nusprendėme pašalinti visus įrašus, kurių Pearsono koreliacijos koeficientas yra didesnis nei |0,69|. Atmesdami sudarėme hipotezę, jog išlaidų (*spendings)* kintamieji suteikia daugiau informacijos už kiekio (*count)* kintamuosius. Atmetus juos liko 19 kintamųjų. Toliau pagal mūsų ekspertų nuomone nusprendėme, kurie kintamieji padės sumodeliuoti tinkamus klasterius ir argumentuotai, kai kuriuos atmetėme:

|  |  |  |
| --- | --- | --- |
| **Kintamieji** | **Apibūdinimas** | **Ar svarbus kintamasis klasterizavimui?** |
| calls\_outgoing\_count | Išeinančių skambučių kiekis | Svarbus, nes imtis tolygiai išsidėsčius ir paprasta interpretuoti. |
| calls\_outgoing\_duration\_max | Ilgiausias skambutis per periodą | Norint sudaryti klasterius šis kintamasis nesukels interpretacijų, kurie svarbus mūsų užduočiai. |
| calls\_outgoing\_inactive\_days | Dienų, be išeinančių skambučių, kiekis | Svarbus, nes parodo, kurie klientai tikėtinai naudojasi kitomis paslaugomis, arba ketina pereiti pas kitą operatorių. |
| calls\_outgoing\_spendings | Išlaidos išeinantiems skambučiams | Svarbus nes gali skirstyti klientus, kurie daugiau naudojasi paslaugomis ir kurie mažiau. |
| calls\_outgoing\_to\_abroad\_count | Skambučių į kitas šalis kiekis | Šių kintamųjų nenaudosime, norint išvengti mažos imties klasterių. 80 procentų klientų net neskambina į užsienį. |
| calls\_outgoing\_to\_abroad\_spendings | Išlaidos skambučiams į užsienį |
| calls\_outgoing\_to\_onnet\_count | Skambučių, tinklo viduje, kiekis | Visa imtis pasiskirsčius ties 0 ir tik 95 procentilis yra 2. Dėl šių priežasčių nenaudosime šio kintamojo. |
| gprs\_inactive\_days | Dienų be duomenų naudojimo kiekis | Nenaudosime dėl to, kad labai mažas kiekis naudoja mobiliuos duomenis ir tai gali sudaryti nedidelės imties klasterių. |
| gprs\_usage | Išnaudotų kb kiekis |
| reloads\_count | Papildymų kiekis | Šie kintamieji gali padėt interpretuoti, kurie klientai pastoviai naudojasi paslaugomis. |
| reloads\_inactive\_days | Dienų, be papildymų, skaičius |
| sms\_outgoing\_inactive\_days | Dienų, be išeinančių SMS žinučių, kiekis | Šie kintamieji nepadės interpretuoti klasterių dėl to kad žinučių kiekis yra labai mažas ir šia funkcija vartotojai beveik nesinaudojo. |
| sms\_outgoing\_spendings | Išlaidos išeinančios SMS žinutėms |
| sms\_outgoing\_spendings\_max | Brangiausia SMS žinutė |
| sms\_outgoing\_to\_abroad\_spendings | Išlaidos išeinančioms SMS žinutėms į kitas šalis |
| sms\_outgoing\_to\_onnet\_count | Išeinančių SMS žinučių tinklo viduje kiekis |
| user\_account\_balance\_last | Kliento sąskaitos likutis periodo gale | Svarbus, nes imtis tolygiai išsidėsčius ir paprasta interpretuoti. |
| user\_lifetime | Kliento buvimo laikas dienomis | Svarbus, nes imtis tolygiai išsidėsčius ir paprasta interpretuoti. |
| user\_no\_outgoing\_activity\_in\_days | Dienų, kai vartotojas neatliko jokių operacijų, kiekis | Svarbus, nes parodo, kurie klientai tikėtinai naudojasi kitomis paslaugomis, arba ketina pereiti pas kitą operatorių. |

Palikome 8 kintamuosius iš usage duomenų, kuriuos naudosime klasterizavimui ir prie jų po agregavimo pridedame churn kintamąjį:

|  |
| --- |
| **Kintamieji** |
| calls\_outgoing\_count |
| calls\_outgoing\_inactive\_days |
| calls\_outgoing\_spendings |
| reloads\_count |
| reloads\_inactive\_days |
| user\_account\_balance\_last |
| user\_lifetime |
| churn |
| user\_no\_outgoing\_activity\_in\_days |

## K-vidurkių algoritmas

Duomenų klasterizavimui buvo pasirinktas K-vidurkių algoritmas. Klasterių (centroidų) skaičiui nustatyti buvo taikomas iteracinis algoritmas, kurio metu visos aktualios metrikos skaičiuojamos iteruojant per skirtingą klasterių skaičių. Optimaliam klasterių skaičių nustatymui buvo panaudota SSE bei f(K) metrikos, kurių priklausomybė nuo klasterių skaičiaus pateikiama Pav. 1 bei Pav. 2.

|  |  |
| --- | --- |
|  |  |
| **Pav. 1.** SSE metrikos priklausomybė nuo klasterių skaičiaus | **Pav. 2. f**(K) metrikos priklausomybė nuo klasterių skaičiaus |

Gautos metrikų priklausomybės rodo, kad optimaliausias sprendinys (klasterių skaičius) yra 5. Tokia išvada daroma dėl to, kad f(K) šioje reikšmėje yra mažiausia, o SSE pagal alkūnės principą įgauna kritinę vertę (lūžio taškas).

## Klasterių interpretavimas

Klasterizavimo metu gauti 5 klasteriai, jų pasiskirstymas pateikiamas Lentelė 1. Iš pateiktų pagal dydį galime išskirti kelias klasterių grupes:

* Klasteris Nr. 1 bei Nr. 3 yra didžiausieji ir sumoje sudaro 74.86 % nuo visų klientų.
* Klasteris Nr. 2 bei Nr. 4 yra vidutinio dydžio ir sumoje sudaro 23,50 % nuo visų klientų.
* Klasteris Nr. 5 yra mažo dydžio ir sudaro likusius 1.59 %.

Lentelė . Klientų pasiskirstymas klasteriuose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Klasteris** | **1** | **2** | **3** | **4** | **5** |
| Klientų skaičius | 18499 | 5485 | 10553 | 3652 | 617 |
| Procentinė dalis | 0.476705 | 0.141344 | 0.271942 | 0.094109 | 0.0159 |
| user\_lifetime | 1810.728 | 15914.72 | 2392.291 | 1168.266 | 9236.828 |
| user\_no\_outgoing\_activity\_in\_days | 3.136233 | 8.350319 | 10.8409 | 1.438299 | 430.2604 |
| user\_account\_balance\_last | 10.02198 | 0.322341 | 7.515656 | 7.854073 | 4.725575 |
| reloads\_inactive\_days | 104.3892 | 34.99654 | 1142.28 | 21.04769 | 858.3144 |
| reloads\_count | 0.720291 | 0.950927 | 0.044632 | 2.240827 | 0.357645 |
| calls\_outgoing\_count | 28.87259 | 0.888453 | 9.803468 | 116.1668 | 16.1718 |
| calls\_outgoing\_spendings | 3.42236 | 0.054646 | 0.678348 | 15.52632 | 1.939941 |
| calls\_outgoing\_inactive\_days | 8.317891 | 1118.935 | 25.07702 | 2.775831 | 442.5667 |
| churn | 0.040056 | 0.863446 | 0.266938 | 0.024096 | 0.444084 |

* Klasteris Nr. 1:

Pagal dydį tai yra didžiausias klasteris kuris nuo visų duomenų sudaro 47,6 %. Šį klasterį sudaro gana aktyvūs bei jauni vartotojai kurių vartojimo statistika yra vidutinė kitų klasterių kontekste. Pastebima, kad šiame klasteryje mažiausias išėjusių klientų skaičius.

* Klasteris Nr. 2:

Pagal dydį tai vidutinis klasteris, kuris nuo visų duomenų sudaro 14,1 %. Šis klasteris išsiskiria tuo, kad jį sudaro seniausi, sąskaitą stabiliai pildantys, bet paslaugomis mažai besinaudojantys klientai. Pastebima, kad šių vartotojų sąskaitos likutis yra mažiausias, o nutraukusių sutartį procentinė dalis kitų klasterių kontekste yra didžiausia. Šiame klasteryje vartotojai ilgiausiai neatliko skambučių, todėl jų suminis skambučių skaičius yra mažiausias.

* Klasteris Nr. 3:

Pagal dydį tai didelis klasteris, kuris nuo visų duomenų sudaro 27,2 %. Šį klasterį sudaro gana jauni ir aktyvūs vartotojai. Labiausiai iš kitų klasterių išsiskiria ilgiausiu sąskaitos nepapildymo laikotarpiu ir dažnumu. Ketvirtadalis šio klasterio klientų nutraukė sutartį.

* Klasteris Nr. 4:

Pagal dydį tai vidutinis klasteris, kuris nuo visų duomenų sudaro 9,4 %. Šį klasteri sudaro jauniausi ir aktyviausi vartotojai. Pastebima, kad šie klientai atlieka daugiausiai skambučių ir turi mažiausius neaktyvumo rodiklius. Taip pat svarbu paminėti, kad būtent šio klasterio mažiausia klientų dalis nutraukė sutartį.

* Klasteris Nr. 5:

Pagal dydį tai yra mažiausias klasteris, kuris nuo visų duomenų sudaro tik 1,6 %. Šį klasterį sudaro gana seni vartotojai kurie išsiskiria savo dideliais neaktyvumo rodikliais. Būtent šio klasterio klientai ilgiausiai neatliko skambučių, nepildė sąskaitos ir nesinaudojo kitomis paslaugomis. Beveik pusė šių klientų nutraukė sutartį.

# Programos kodas

Programos kodas . Duomenų agregavimas

|  |
| --- |
| **import** logging  **import** os  **import** glob  **import** pandas **as** pd  **from** pathlib **import** Path  **from** dotenv **import** find\_dotenv**,** load\_dotenv  **import** shutil  **import** click  **import** pyspark  LOG\_FORMAT **=** **(**  "%(asctime)s.%(msecs)03d - %(levelname)s - "  "%(filename)s - %(lineno)s - %(message)s"**)**  logging**.**basicConfig**(**  level**=**logging**.**INFO**,**  format**=**LOG\_FORMAT**,**  datefmt**=**"%Y-%m-%d,%H:%M:%S"**)**  TMP\_DIR **=** 'aggregated\_customer\_usage\_TMP'  *@click***.**command**()**  *@click***.**argument**(**'path\_usage'**,** type**=**click**.**Path**(**exists**=True))**  *@click***.**argument**(**'path\_output'**,** type**=**click**.**Path**())**  **def** main**(**path\_usage**,** path\_output**):**  logger **=** logging**.**getLogger**(**\_\_name\_\_**)**  tmp\_dir\_path **=** os**.**path**.**join**(**path\_output**,** TMP\_DIR**)**  spark **=** **(**  pyspark**.**sql**.**SparkSession  **.**builder  **.**appName**(**"Python Spark SQL aggregation example"**)**  **.**enableHiveSupport**()**  **.**getOrCreate**()**  **)**  os**.**makedirs**(**path\_output**,** exist\_ok**=True)**  usage\_df **=** spark**.**read**.**csv**(**path\_usage**,** header**=True,** inferSchema**=True,** sep**=**','**)**  usage\_df**.**createOrReplaceTempView**(**"customer\_usage"**)**  date\_columns **=** **[**"year"**,** "month"**]**  id\_columns **=** **[**"user\_account\_id"**]**  binary\_columns **=** **[**  "user\_intake"**,**  "user\_has\_outgoing\_calls"**,** "user\_has\_outgoing\_sms"**,**  "user\_use\_gprs"**,** "user\_does\_reload"  **]**  categorical\_columns **=** date\_columns **+** binary\_columns **+** id\_columns  continuous\_columns **=** **[**c **for** c **in** usage\_df**.**columns **if** c **not** **in** categorical\_columns**]**  sql\_expressions\_avg **=** **[**"AVG({0}) AS {0}"**.**format**(**c**)** **for** c **in** continuous\_columns**]**  sql\_expressions\_max **=** **[**"MAX({0}) AS {0}"**.**format**(**c**)** **for** c **in** binary\_columns**]**  sql\_expressions\_count **=** **[**"COUNT(\*) AS n\_months"**]**  sql\_expressions\_aggregation **=** sql\_expressions\_avg **+** sql\_expressions\_max **+** sql\_expressions\_count  sql\_query\_aggregate\_by\_user\_id **=** """  SELECT user\_account\_id, {}  FROM customer\_usage  GROUP BY user\_account\_id  """**.**format**(**"\n , "**.**join**(**sql\_expressions\_aggregation**))**  aggregate\_usage\_df **=** spark**.**sql**(**sql\_query\_aggregate\_by\_user\_id**)**  aggregate\_usage\_df**.**write**.**csv**(**tmp\_dir\_path**)**  **with** open**(**os**.**path**.**join**(**path\_output**,** TMP\_DIR**,** "header\_\_aggregated\_customer\_usage.txt"**),** "w"**)** **as** f**:**  f**.**write**(**","**.**join**(**aggregate\_usage\_df**.**columns**)** **+** "\n"**)**  **with** open**(**os**.**path**.**join**(**path\_output**,** TMP\_DIR**,** "header\_\_aggregated\_customer\_usage.txt"**))** **as** f**:**  columns **=** f**.**read**().**rstrip**().**split**(**","**)**  combined\_df **=** join\_csv\_files**(**tmp\_dir\_path**,** columns**)**  combined\_df**.**to\_csv**(**os**.**path**.**join**(**tmp\_dir\_path**,** ".."**,** "aggregated\_customer\_usage.csv"**),** index**=False,**  encoding**=**'utf-8-sig'**)**  **try:**  shutil**.**rmtree**(**tmp\_dir\_path**)**  **except** OSError **as** e**:**  **print(**"Error: %s - %s." **%** **(**e**.**filename**,** e**.**strerror**))**  **def** join\_csv\_files**(**file\_dir**,** coll\_names**):**  extension **=** 'csv'  all\_filenames **=** **[**i **for** i **in** glob**.**glob**(**file\_dir **+** '/\*.{}'**.**format**(**extension**))]**  combined\_csv **=** pd**.**concat**([**pd**.**read\_csv**(**f**,** sep**=**','**,** names**=**coll\_names**)** **for** f **in** all\_filenames**])**  **return** combined\_csv  **if** \_\_name\_\_ **==** '\_\_main\_\_'**:**  log\_fmt **=** '%(asctime)s - %(name)s - %(levelname)s - %(message)s'  logging**.**basicConfig**(**level**=**logging**.**INFO**,** format**=**log\_fmt**)**  # not used in this stub but often useful for finding various files  project\_dir **=** Path**(**\_\_file\_\_**).**resolve**().**parents**[**2**]**  # find .env automagically by walking up directories until it's found, then  # load up the .env entries as environment variables  load\_dotenv**(**find\_dotenv**())**  main**()** |

Programos kodas . Duomenų sujungimas

|  |
| --- |
| **import** logging  **from** pathlib **import** Path  **from** dotenv **import** find\_dotenv**,** load\_dotenv  **import** click  **import** pyspark  LOG\_FORMAT **=** **(**  "%(asctime)s.%(msecs)03d - %(levelname)s - "  "%(filename)s - %(lineno)s - %(message)s"**)**  logging**.**basicConfig**(**  level**=**logging**.**INFO**,**  format**=**LOG\_FORMAT**,**  datefmt**=**"%Y-%m-%d,%H:%M:%S"**)**  *@click***.**command**()**  *@click***.**argument**(**'path\_usage\_csv'**,** type**=**click**.**Path**(**exists**=True))**  *@click***.**argument**(**'path\_churn\_csv'**,** type**=**click**.**Path**(**exists**=True))**  *@click***.**argument**(**'path\_output'**,** type**=**click**.**Path**())**  **def** main**(**path\_usage\_csv**,** path\_churn\_csv**,** path\_output**):**  logger **=** logging**.**getLogger**(**\_\_name\_\_**)**  spark **=** **(**  pyspark**.**sql**.**SparkSession  **.**builder  **.**appName**(**"Python Spark SQL aggregation with join"**)**  **.**enableHiveSupport**()**  **.**getOrCreate**()**  **)**  usage\_df **=** spark**.**read**.**csv**(**  path\_usage\_csv**,**  header=True,  inferSchema=True)  churn\_df = spark.read.csv(  path\_churn\_csv,  header=True,  inferSchema=True)  agg\_usage\_churn\_df = (  usage\_df  .join(churn\_df, "user\_account\_id")  )  agg\_usage\_churn\_df.write.parquet(path\_output)  if \_\_name\_\_ == '\_\_main\_\_':  log\_fmt = '%(asctime)s - %(name)s - %(levelname)s - %(message)s'  logging.basicConfig(level=logging.INFO, format=log\_fmt)  # not used in this stub but often useful for finding various files  project\_dir = Path(\_\_file\_\_).resolve().parents[2]  # find .env automagically by walking up directories until it's found, then  # load up the .env entries as environment variables  load\_dotenv(find\_dotenv())  main() |

Programos kodas . Duomenų klasterizavimas

|  |
| --- |
| **import** logging  **from** pathlib **import** Path  **from** dotenv **import** find\_dotenv**,** load\_dotenv  **import** click  **import** os  **import** operator  **import** jsonlines  **import** pandas **as** pd  **import** pyspark  **from** pyspark**.**ml**.**feature **import** VectorAssembler**,** StandardScaler  **from** pyspark**.**ml**.**pipeline **import** Pipeline  **from** pyspark**.**ml**.**clustering **import** KMeans**,** GaussianMixture  **import** json  **import** pyspark  LOG\_FORMAT **=** **(**  "%(asctime)s.%(msecs)03d - %(levelname)s - "  "%(filename)s - %(lineno)s - %(message)s"**)**  logging**.**basicConfig**(**  level**=**logging**.**INFO**,**  format**=**LOG\_FORMAT**,**  datefmt**=**"%Y-%m-%d,%H:%M:%S"**)**  *@click***.**command**()**  *@click***.**argument**(**'params\_json'**,** type**=**click**.**Path**(**exists**=True))**  **def** main**(**params\_json**):**  logger **=** logging**.**getLogger**(**\_\_name\_\_**)**  **with** open**(**params\_json**)** **as** json\_file**:**  settings\_json **=** json**.**load**(**json\_file**)**  spark **=** **(**  pyspark**.**sql**.**SparkSession  **.**builder  **.**appName**(**"Python Spark K-means"**)**  **.**enableHiveSupport**()**  **.**getOrCreate**()**  **)**  path\_aggregated\_df **=** settings\_json**[**'path\_aggregated\_df'**]**  clustering\_df **=** spark**.**read**.**parquet**(**path\_aggregated\_df**)**  columns\_clustering\_features **=** columns\_clustering\_features **=** **[**  'user\_lifetime'**,**  'user\_no\_outgoing\_activity\_in\_days'**,**  'user\_account\_balance\_last'**,**  'user\_spendings'**,**  'reloads\_inactive\_days'**,**  'reloads\_count'**,**  'calls\_outgoing\_count'**,**  'calls\_outgoing\_spendings\_max'**,**  'calls\_outgoing\_inactive\_days'**,**  'calls\_outgoing\_to\_onnet\_count'**,**  'calls\_outgoing\_to\_onnet\_spendings'**,**  'calls\_outgoing\_to\_abroad\_count'**,**  'calls\_outgoing\_to\_abroad\_duration'**,**  'sms\_outgoing\_count'**,**  'sms\_outgoing\_spendings\_max'**,**  'sms\_outgoing\_inactive\_days'**,**  'sms\_outgoing\_to\_onnet\_count'**,**  'sms\_outgoing\_to\_abroad\_count'**,**  'gprs\_session\_count'**,**  'gprs\_spendings'**,**  'gprs\_inactive\_days'**,**  **]**  vector\_assembler **=** VectorAssembler**(**  inputCols**=**columns\_clustering\_features**,**  outputCol**=**"initial\_features"**)**  standard\_scaler **=** StandardScaler**(**  inputCol**=**"initial\_features"**,**  outputCol**=**"features"**,**  withStd**=True,**  withMean**=True)**  vectorized\_df **=** vector\_assembler**.**transform**(**clustering\_df**)**  model\_scaler **=** standard\_scaler**.**fit**(**vectorized\_df**)**  featurized\_clustering\_df **=** model\_scaler**.**transform**(**vectorized\_df**)**  featurization\_pipeline **=** Pipeline**(**stages**=[**vector\_assembler**,** standard\_scaler**])**  featurization\_pipeline\_model **=** featurization\_pipeline**.**fit**(**clustering\_df**)**  model\_scaler **=** featurization\_pipeline\_model**.**stages**[-**1**]**  featurized\_clustering\_df **=** featurization\_pipeline\_model**.**transform**(**clustering\_df**)**  **for** k **in** settings\_json**[**'k\_values'**]:**  kmeans **=** KMeans**(**featuresCol**=**"features"**,** k**=**k**)**  model\_kmeans **=** kmeans**.**fit**(**featurized\_clustering\_df**)**  path\_metrics\_kmeans\_sse **=** settings\_json**[**'path\_metrics\_kmeans\_sse'**]**  sse **=** model\_kmeans**.**computeCost**(**featurized\_clustering\_df**)**  metrics\_row **=** **{**"k"**:** k**,** "sse"**:** sse**}**  **with** jsonlines**.**open**(**path\_metrics\_kmeans\_sse**,** "a"**)** **as** f**:**  f**.**write**(**metrics\_row**)**  normalized\_cluster\_centers **=** model\_kmeans**.**clusterCenters**()**  scaler\_mean **=** model\_scaler**.**mean  scaler\_std **=** model\_scaler**.**std  cluster\_sizes **=** model\_kmeans**.**summary**.**clusterSizes  n\_obs **=** clustering\_df**.**count**()**  denormalized\_cluster\_centers **=** **[**  **(**cluster\_id**,)** **+** **(**size**,** 100 **\*** size **/** n\_obs**)** **+** tuple**(**center **\*** scaler\_std **+** scaler\_mean**)**  **for** cluster\_id**,** **(**size**,** center**)** **in**  enumerate**(**zip**(**cluster\_sizes**,** normalized\_cluster\_centers**))**  **]**  cluster\_centers\_pddf **=** pd**.**DataFrame**.**from\_records**(**denormalized\_cluster\_centers**)**  cluster\_centers\_pddf**.**columns **=** **(**  **[**"cluster\_id"**,** "cluster\_size"**,** "cluster\_size\_pct"**]** **+**  columns\_clustering\_features  **)**  pd**.**set\_option**(**"max\_columns"**,** 999**)**  path\_cluster\_centers **=** settings\_json**[**'path\_cluster\_centers'**]** **+** "cluster\_centers\_kmeans\_\_k\_{}.csv"**.**format**(**k**)**  cluster\_centers\_pddf**.**to\_csv**(**path\_cluster\_centers**,** index**=False)**  clustered\_kmeans\_df **=** model\_kmeans**.**transform**(**featurized\_clustering\_df**)**  path\_clustered\_df **=** settings\_json**[**'path\_cluster\_centers'**]** **+** "clustered\_kmeans\_\_k\_{}\_parquet"**.**format**(**k**)**  clustered\_kmeans\_df**.**write**.**parquet**(**path\_clustered\_df**)**  **if** \_\_name\_\_ **==** '\_\_main\_\_'**:**  log\_fmt **=** '%(asctime)s - %(name)s - %(levelname)s - %(message)s'  logging**.**basicConfig**(**level**=**logging**.**INFO**,** format**=**log\_fmt**)**  # not used in this stub but often useful for finding various files  project\_dir **=** Path**(**\_\_file\_\_**).**resolve**().**parents**[**2**]**  # find .env automagically by walking up directories until it's found, then  # load up the .env entries as environment variables  load\_dotenv**(**find\_dotenv**())**  main**()** |

Programos kodas . f(K) metrikos apskaičiavimas

|  |
| --- |
| #!/usr/bin/env python  # coding: utf-8  # # $f(K)$ metrikos naudojimo pavyzdys K-vidurkių klasterizavimo rezultatams  # In[10]:  get\_ipython**().**run\_line\_magic**(**'load\_ext'**,** 'autoreload'**)**  get\_ipython**().**run\_line\_magic**(**'autoreload'**,** '2'**)**  get\_ipython**().**run\_line\_magic**(**'matplotlib'**,** 'inline'**)**  # In[11]:  **from** IPython**.**display **import** Image  **import** plotnine **as** gg  **import** pandas **as** pd  # `pyspark.ml.clustering.KMeansModel` metodas `computeCost` apskaičiuoja stebėjimų Euklido atstumų nuo savo klasterių centrų sumą $S\_K$ (angl. \_Within Set Sum of Squared Error (WSSSE)\_):  #  # $I\_k = \sum\_{\mathbf{x}\_i \in C\_k} \| \mathbf{x}\_i - \mathbf{\overline{x}}\_k \|$  #  # $S\_K = \sum\_{k}^{K} I\_k$  #  # čia  #  # $k$ - klasterio indeksas,  #  # $C\_k$ - $k$-asis klasteris  #  # $K$ - klasterių skaičius,  #  # $N\_k$ - $k$-jam klasteriui priklausančių stebėjimų skaičius,  #  # $\mathbf{x\_i}$ - $i$-tojo stebėjimo vektorius,  #  # $\mathbf{\overline{x}}\_k$ - $k$-otojo klasterio vidurinio taško (centro) vektorius,  #  # $\|\mathbf{x}\|$ - vektoriaus Euklido norma, t.y. kvadratinė šaknis iš jo komponenčių kvadratų sumos.  # $f(K)$ yra naudojama nustatyti optimalią $K$ reikšmę ir yra aprašyta [čia](http://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf) ir [čia](https://datasciencelab.wordpress.com/2014/01/21/selection-of-k-in-k-means-clustering-reloaded/). Jos reikšmė $f(K)$ apskaičiuojama tokiu būdu:  # In[12]:  # Image("../data/img/fk.png")  # Optimilaus $K$ yra ties mažiausia $f(K)$ reikšme.  # Realizuojame formulės išraišką.  # In[13]:  **def** compute\_fk**(**k**,** sse**,** prev\_sse**,** dim**):**  **if** k **==** 1 **or** prev\_sse **==** 0**:**  **return** 1  weight **=** weight\_factor**(**k**,** dim**)**  **return** sse **/** **(**weight **\*** prev\_sse**)**  # calculating alpha\_k in functional style with tail recursion -- which is not optimized in Python :(  **def** weight\_factor**(**k**,** dim**):**  **if** **not** k **>** 1**:**  **raise** ValueError**(**"k must be greater than 1"**)**    **def** weigth\_factor\_accumulator**(**acc**,** k**):**  **if** k **==** 2**:**  **return** acc  **return** weigth\_factor\_accumulator**(**acc **+** **(**1 **-** acc**)** **/** 6**,** k **-** 1**)**    weight\_k2 **=** 1 **-** 3 **/** **(**4 **\*** dim**)**  **return** weigth\_factor\_accumulator**(**weight\_k2**,** k**)**  # Aprašome funkciją, kuri iš $K$ ir $S\_K$ reikšmių porų `list`'o pateikia galimas įvertinti $f(K)$ reikšmes.  # In[14]:  **def** compute\_fk\_from\_k\_sse\_pairs**(**k\_sse\_pairs**,** dimension**):**  triples **=** make\_fk\_triples**(**k\_sse\_pairs**)**  k\_fk\_pairs **=** **[**  **(**k**,** compute\_fk**(**k**,** sse**,** prev\_sse**,** dimension**))**  **for** **(**k**,** sse**,** prev\_sse**)** **in** triples**]**  **return** sorted**(**k\_fk\_pairs**,** key**=lambda** pair**:** pair**[**0**])**  **def** make\_fk\_triples**(**k\_sse\_pairs**):**  sorted\_pairs **=** sorted**(**k\_sse\_pairs**,** reverse**=True)**  candidates **=** list**(**zip**(**sorted\_pairs**,** sorted\_pairs**[**1**:]** **+** **[(**0**,** 0.0**)]))**  triples **=** **[**  **(**k**,** sse**,** prev\_sse**)**  **for** **((**k**,** sse**),** **(**prev\_k**,** prev\_sse**))** **in** candidates  **if** k **-** prev\_k **==** 1  **]**  **return** triples  # Naudojimo pavyzdys:  # In[15]:  get\_ipython**().**system**(**' cat ../data/examples\_io/metrics\_\_k\_means\_\_sse.jsonl'**)**  # Tarkime, iš disko nuskaitome tokią $K$ ir $S\_K$ reikšmių lentelę.  # In[16]:  metrics\_pddf **=** pd**.**read\_json**(**  "../data/examples\_io/metrics\_\_k\_means\_\_sse.jsonl"**,**  orient**=**"records"**,**  lines**=True)**  metrics\_pddf  # Pakeičiame stulpelių tvarką.  # In[17]:  k\_sse\_pddf **=** metrics\_pddf**[[**"k"**,** "sse"**]]**  k\_sse\_pddf  # Skačiuojant $f(K)$ metriką reikia žinoti duomenų dimensiją, t.y. klasterizavimui naudotų požymių skaičių. Tarkime, kad šiuo atveju naudojome du požymius.  # In[18]:  dimension **=** 2  k\_sse\_pairs **=** **[**tuple**(**r**)** **for** r **in** k\_sse\_pddf**.**to\_records**(**index**=False)]**  k\_sse\_pairs  # In[19]:  k\_fk\_pairs **=** compute\_fk\_from\_k\_sse\_pairs**(**k\_sse\_pairs**,** dimension**)**  k\_fk\_pairs  # In[20]:  k\_fk\_pddf **=** pd**.**DataFrame**.**from\_records**(**k\_fk\_pairs**,** columns**=[**"k"**,** "fk"**])**  k\_fk\_pddf  # In[21]:  plot\_k\_sse **=** **(**  gg**.**ggplot**(**gg**.**aes**(**x**=**"k"**,** y**=**"sse"**),** data**=**k\_sse\_pddf**)** **+**  gg**.**geom\_line**()** **+**  gg**.**xlab**(**"K"**)** **+**  gg**.**ylab**(**"SSE"**)** **+**  gg**.**ggtitle**(**"SSE pagal klasterių skaičių K"**)** **+**  gg**.**theme\_bw**()**  **)**  plot\_k\_fk **=** **(**  gg**.**ggplot**(**gg**.**aes**(**x**=**"k"**,** y**=**"fk"**),** data**=**k\_fk\_pddf**)** **+**  gg**.**geom\_line**()** **+**  gg**.**xlab**(**"K"**)** **+**  gg**.**ylab**(**"f(K)"**)** **+**  gg**.**ggtitle**(**"f(K) pagal klasterių skaičių K"**)** **+**  gg**.**theme\_bw**()**  **)**  # In[22]:  **print(**plot\_k\_sse**)**  **print(**plot\_k\_fk**)** |