KAUNO TECHNOLOGIJOS UNIVERSITETAS
Informatikos fakultetas

Intelektikos pagrindai

Individualus darbas

Data: 2021-05-07

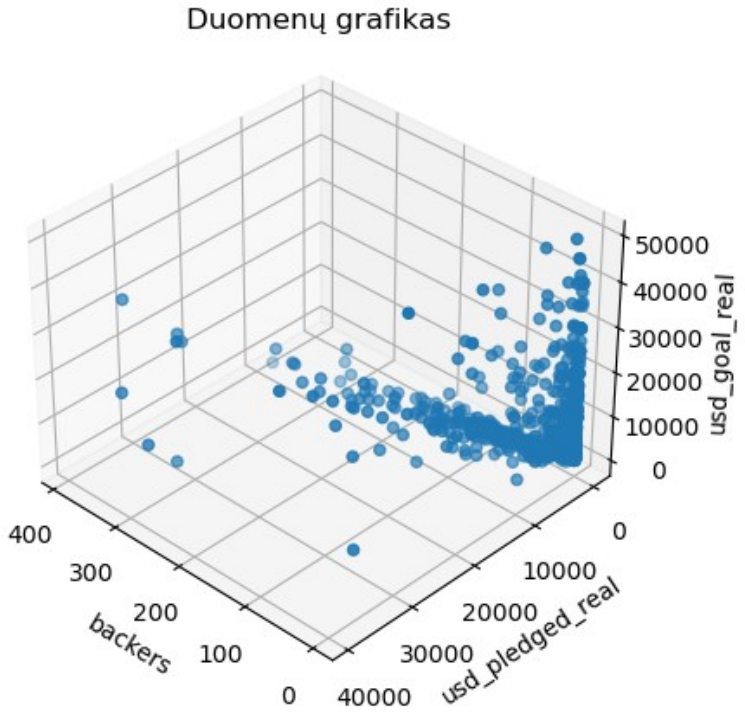
Dėstytojai:
Agnė Paulaiskaitė-Tarasevičienė
Germanas Budnikas
Studentai:
Arvydas Miklovis IFF-8/10
Dainius Čepulis IFF8/10

KAUNAS, 2021

Turinys

1 Darbo dalys....2
2. Įvadas....2
3. Duomenų pasiruošimas....2
4. Gaussian Mixture....2
4.1. Modelio pritaikymas....2
4.2. Išvados....5
5. K-vidurkių metodas....6
5.1. Inercija....6
5.2. Silueto koeficiantas....6
5.3. Vizualizacija....7
5.4. Išvados....9
7. Apibendrintos išvados....10

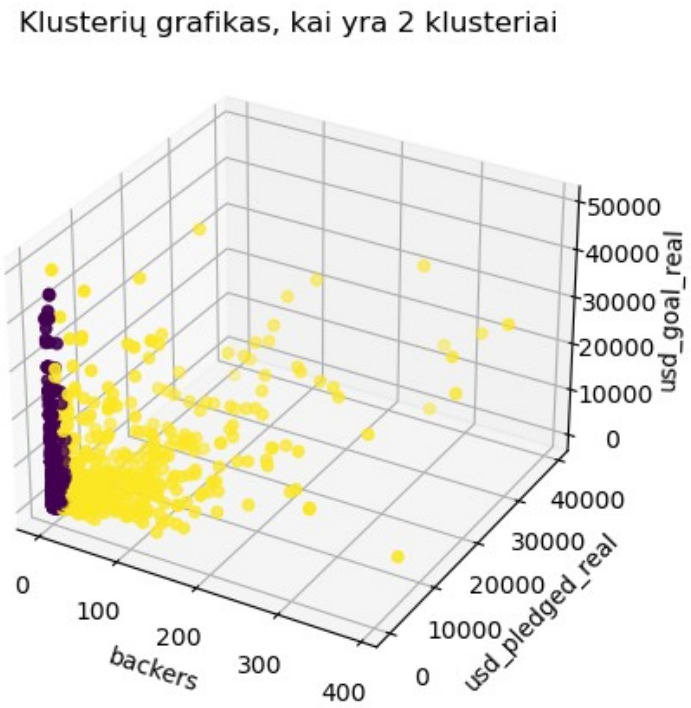
1. Darbo dalys
Gaussian Mixture - Dainius Čepulis
K-vidurkiai- Arvydas Miklovis
2. Įvadas
Metodams palyginti buvo naudotas duomenų rinkinys iš pirmo laboratorinio darbo. Iš jo naudosime tik 3 atributus, nes buvo pastebėtai, kad jie vizualiai išsiskirto klasterius. Duomenų rinkinys dėl paprastumo buvo sumažintas, tačiau tai menkai, keičia atsakymus.
Taip pat galime pažiūrėti į pasirinktų duomenų grafika.



Iš grafiko, matome kad vizualiai duomenis galime ganėtinai lengvai padalinti į du klasterius, vienas susitelkęs prie „backers“ ir „usd_pledged_real“ ašies, o kitas yra sudarytas iš likusiu duomenų, kuris sudaro „šluotos“ formos plokštumą.

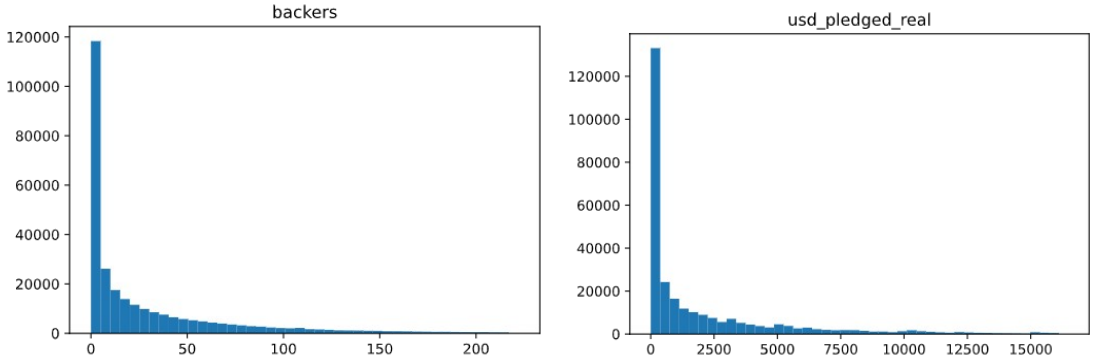
pav. 1 pasirinktų duomenų grafikas

3. Duomenų pasiruošimas
Kadangi mūsų duomenys turėjo įtin ekstremalių reikšmių, prieš taikydami duomenis modeliuose, jas pašalinome.
4. Gaussian Mixture
Kadangi buvo sunku atrast python bibliotekas, kurios palaike SOM su klasterių limitu, buvo pasirinktas kitas metodas.
Gaussian Mixture yra mašininio mokimosi metodas, kuris atranda klasterius darydamas prielaida, kad šie klasteriai sudaro normalų pasiskirstymą.
4.1. Modelio pritaikymas
Kad įvertintumėme, kaip gerai šis metodas iškrito klasterius ir kada jis susidaro geriausius klasterius keisime jų skaičių (2,3,5,9). Ir gautus rezultatus parodysime trimatėje erdvėje, kiekvieną duomenų imties elementą nuspalvinę jam priskirto klasterio spalva.



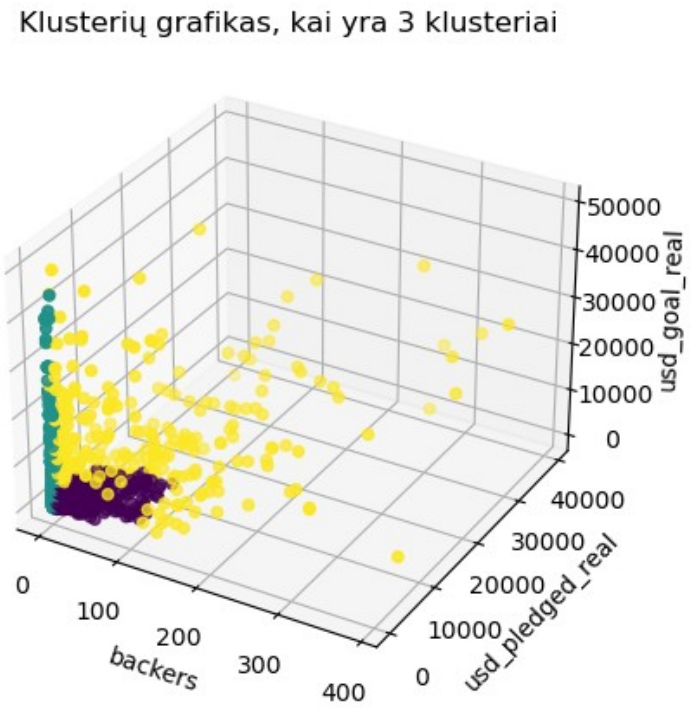
pav. 2. Išskirti du klasteriai

Kai buvo pasirinkti 2 klusteriai matome, duomenis buvo išskirstyti, kurie turėjo mažai „backers“ ir „usd_pledged_real“. Ką tai galimai reikštu galime pažiūrėti iš histogramų.



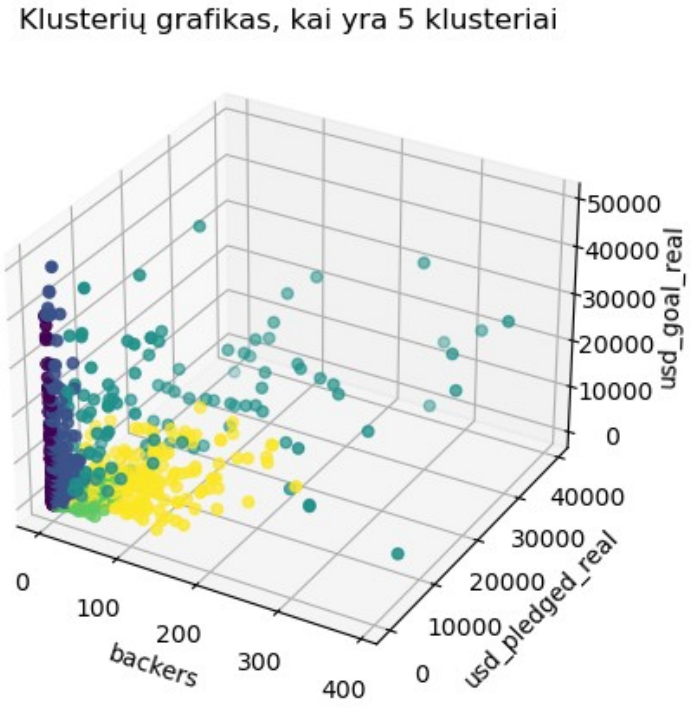
pav. 3. „backers“ ir „usd_pleadged_real“ histogramos

Ir šių duomenų histogramų matome, kad tai atsitiko, kad tokiu reikšmių buvo labai daug. Ir tai yra ganėtinai geras pasirinkimas, nes tai yra galimai nepasisekė projektai.



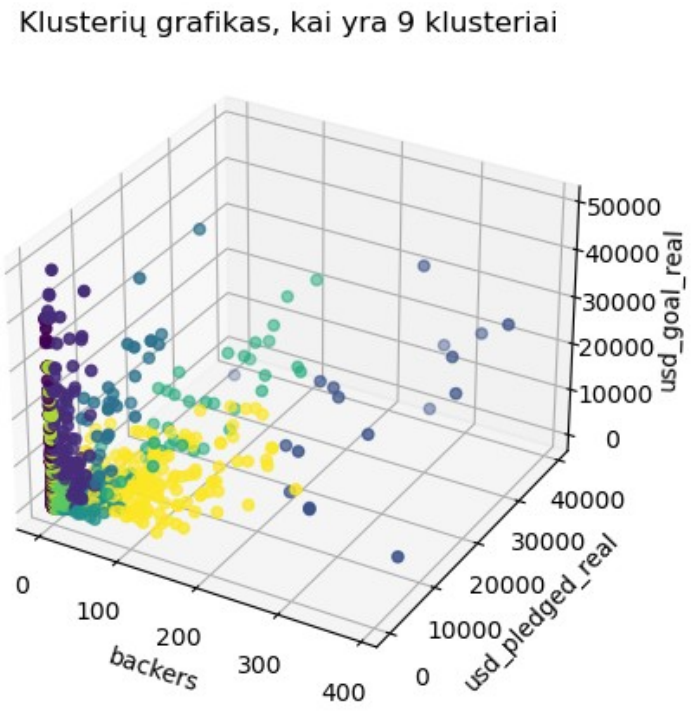
pav. 4. Išskirti trys klasteriai

Kai nustatėme klasterių į 3, mūsų prieš tai aptartas klasteris išliko, o trečias klasteris buvo atskirtas nuo antro. Ir galima spėti, kad buvo atskirti maži, galimai pasisekė projektai.



pav. 5. Išskirti penki klasteriai

Iš 5 klasterių grafiko jau sunku nuspręsti, ką kiekvienas klasteris gali reikšti. Tačiau jie neatrodo pernelyg netikėtini.



pav. 6. Išskirti devini klasteriai

Kai sudarome 9 klasterių grafika, jis pasidaro sunkiai skaitomas ir galime tikėtis, kad pasirinkome per daug klasterių.

4.2. Išvados

Kadangi dažnai raliuose duomenyse yra normaliojo pasiskirstymo tendencijų, šis metodas buvo neblogas pasirinkimas ir išskirti klasteriai, kai pasirenkama tarp 2 ir 5 klasterių, buvo ganėtinai tikėtini, tačiau dėl darbo apimties buvo sunkų tai įvertinti.

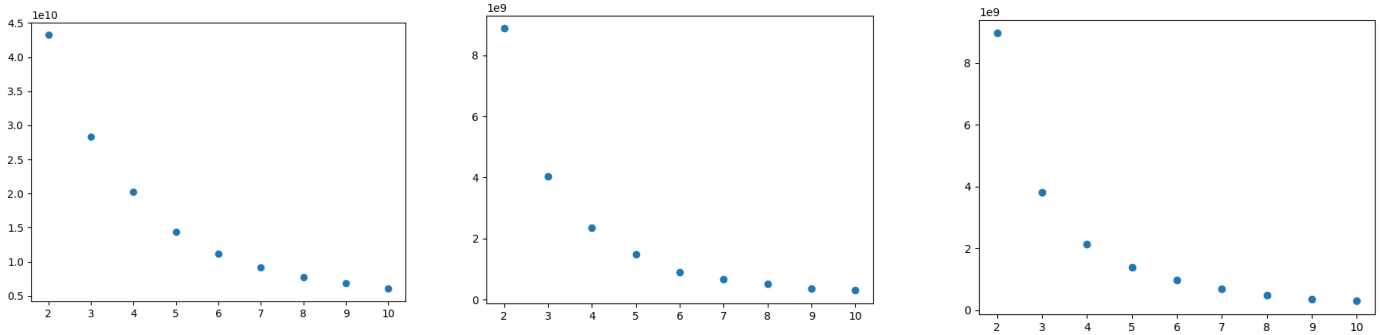
5. K-vidurkių metodas

K-vidurkių metodas – neprižiūrimojo tipo duomenų panašumu grįstas algoritmas, kuris duomenis bando susiskirstyti į *K* nepersidengiančių klasterių. Nustatyti, ar gerai atliktas grupavimas, būtų galima, jeigu turėtume atsakymus; deja, šiuose uždaviniuose tokios informacijos nėra. K-vidurkių metodas labiau skirtas panašumui tarp duomenų nustatyti ir pagal tai pasidaryti tam tikras išvadas. Tam tikrais atvejais duomenys labai aiškiai skiriasi pagal konkrečias savybes, o kartais jie būna per daug panašūs, ir toks sugrupavimas gali būti betikslis.

Modelis skaičiuotas 3 skirtingomis atributų variacijomis
Pirmas variantas: `backers, usd_pledged_real, usd_goal_real`
Antras variantas: `project_time, usd_pledged_real, usd_goal_real`
Trečias variantas: `backers, usd_pledged_real, project_time`

5.1. Inercija

Klasteriai	Inercijos reikšmė NR1	Inercijos reikšmė NR2	Inercijos reikšmė NR3
2	43228073402	8874580746	8968415223
3	28279830243	4042952354	3822256545
4	20201370010	2349934956	2128245239
5	14439257859	1490596688	1396440898
6	11200675896	896960720	975760813
7	9183050286	662409575	682575219
8	7703207724	504637797	493187101
9	6886510287	370483087	359299982
10	6041935061	300983535	295338352

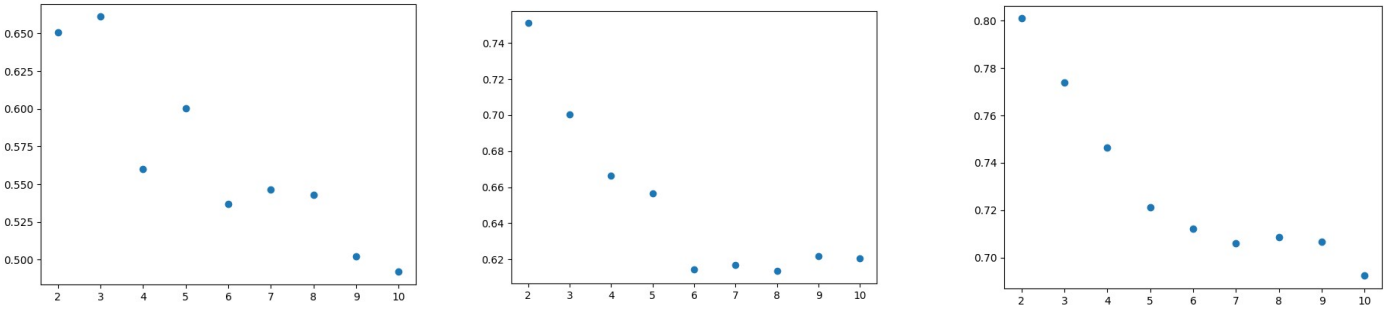


pav. 7. Inercijos ir klasterių priklausomybė

Pagal gautus rezultatus, matome, jog vadinamoji „alkūnė“- rekomenduojamas klasterių kiekis skiriasi nuo atributų variantų. Pirmame variante tinkami pasirinkimai būtų 5-7 klasteriai, kai tuo tarpu antrame ir trečiame variante alkūnė gauname ties 4-6 klasteriais.

5.2. Silueto koeficiantas

Klasteriai	Inercijos reikšmė NR1	Inercijos reikšmė NR2	Inercijos reikšmė NR3
2	0.650	0.751	0.800
3	0.661	0.700	0.773
4	0.559	0.666	0.746
5	0.600	0.656	0.721
6	0.536	0.614	0.712
7	0.546	0.616	0.705
8	0.542	0.613	0.708
9	0.502	0.621	0.706
10	0.491	0.620	0.692

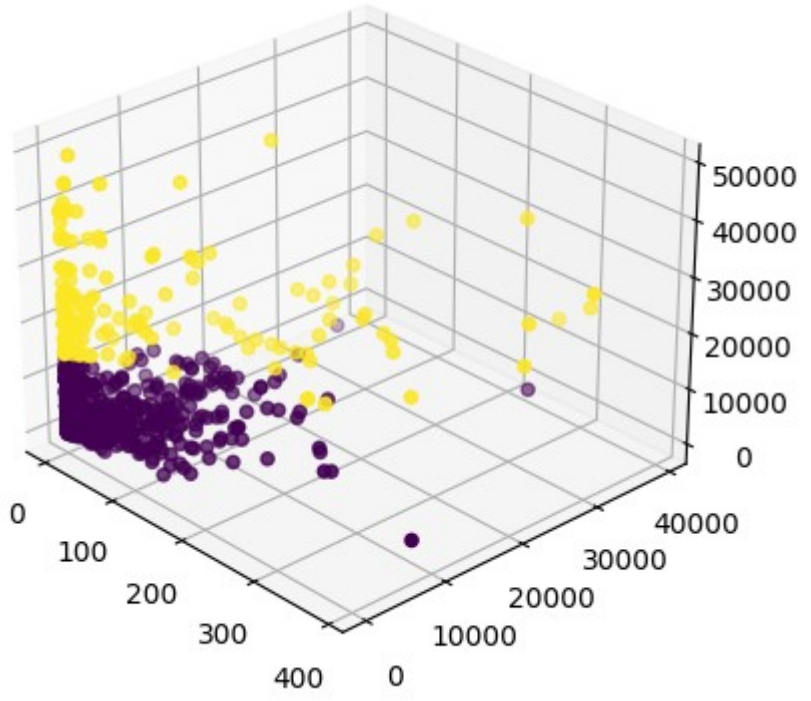


pav. 8. Silueto koeficiantas

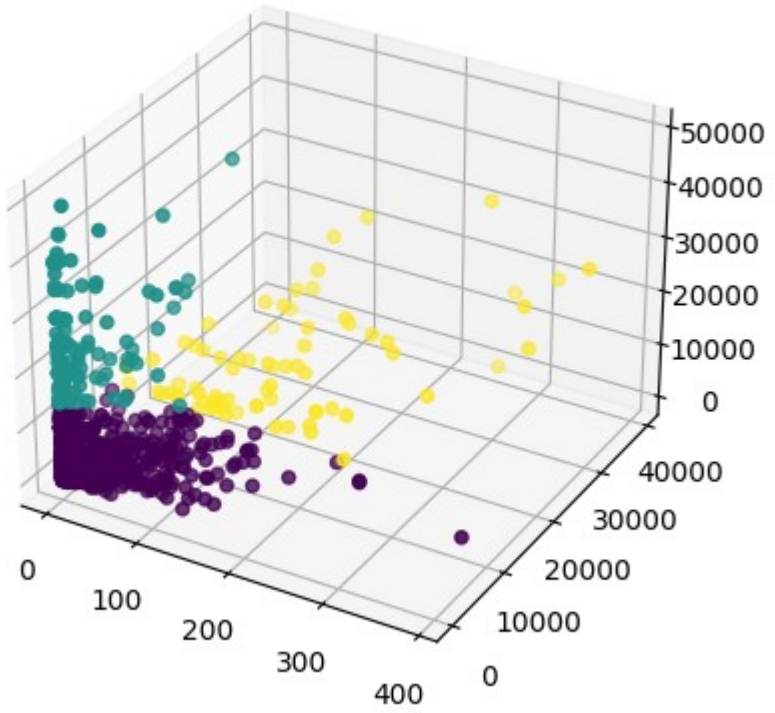
Vėlgi, kaip ir su inercija, matome, kad rezultatai skiriasi pagal atributų variacijas. Pirmuoju variantu gauname aukščiausią įverti ties 3 klasteriais, kai tuo tarpu antrame ir trečiame variantuose, geriausi rezultatai rinkinį skeliant tik į du klasterius. Taip pat verta paminėti, jog labiausiai klasteriai išsiskyrė trečiuoju duomenų rinkinių ir įvertis gaunamas gana aukštas- 0,8.

5.3. Vizualizacija

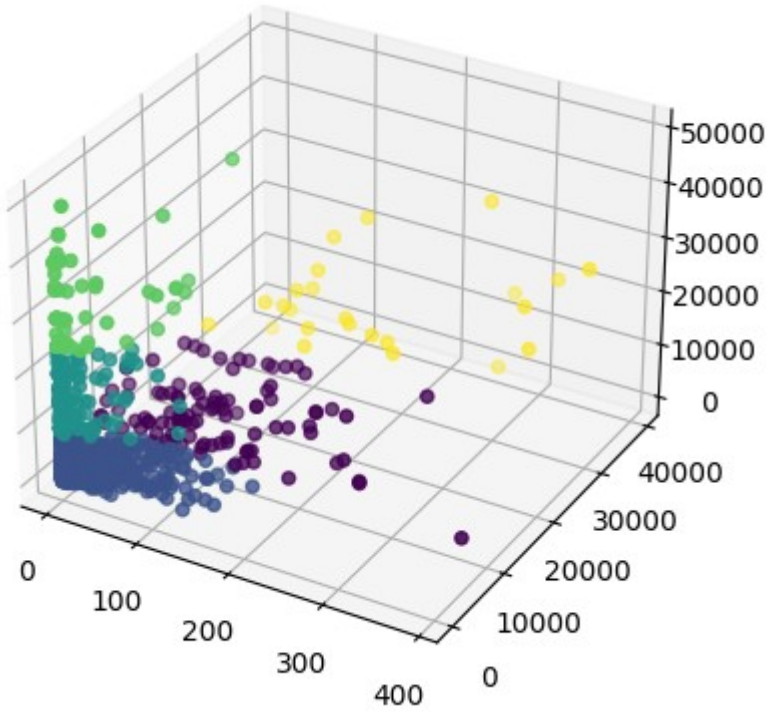
Taip pat pateikiama trečiojo duomenų rinkinio klasterių vizualizacija



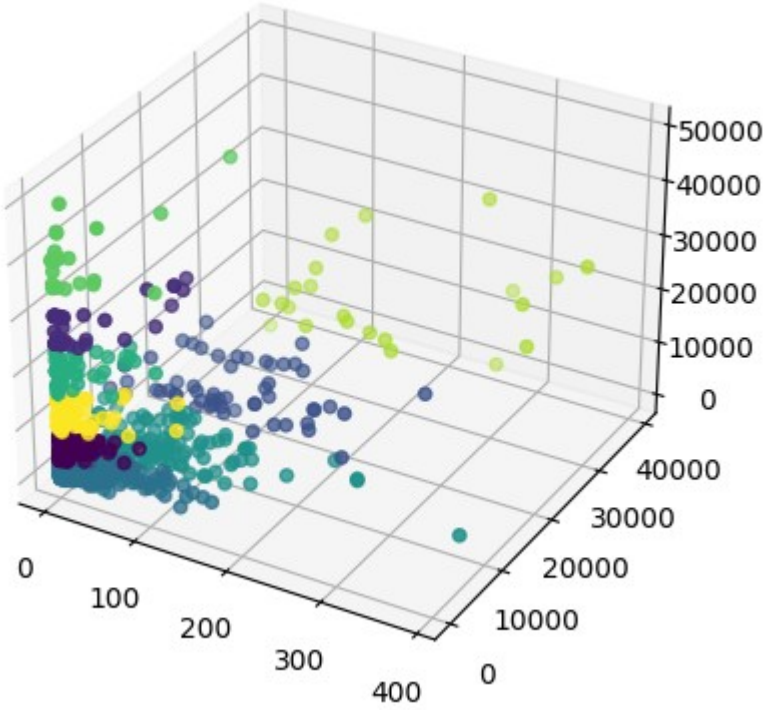
pav. 9. Išskirti du klasteriai



pav. 10. Išskirti trys klasteriai



pav. 11. Išskirti penki klasteriai



pav. 12. Išskirti devyni klasteriai

5.4. Išvados

Priimti vienareikšmį sprendimą šitam moduliui klasterizuoti itin sudėtinga. Inertijos ir silueto metrikos sufleruoja skirtingus atsakymus. Manau sprendžiant tokią problemą taip pat reikėtų diskutuoti ir su užsakovu, galbūt jis įžvelgtu vieno ar kito pasirinkimo pranašumus ir trūkumus, bet klasterizuojant šį duomenų rinkinį rinkčiausi 2-4 klasterius. Silueto metodas sufleruoja, jog rinktis reiktu 2, inertijos alkūnė gaunama ties 4-6 klasteriais, tai 3-4 klasteriai būtų logiškiausias pasirinkimas, juolab, jog rezultatai nesiskiria drastiškai, viskas keičiasi itin palaipsniui.

6. Apibendrintos išvados

Gausian Mixture metodo vizualizacijos rodo, kad šiam modeliui klasterizuoti tinkami 2,3,5 klasteriai, su kiek didesniu- 9 klasteriais, matomas klasterių persidengimas, todėl daugiau nei daugiau nei 5 klasterių nerekomenduojama rinktis sprendžiant šią problemą.

K-vidurkių metodo skaitiniai įverčiai rodo panašius rezultatus, bendras ivertinimas galėtų būti 2-5 klasteriai, nors atskiros metrikos sufleruoja skirtingus rezultatus.

Taip pat verta paminėti, jog iš metodų vizualizacijų išsidėstymo matome, kad klasteriai erdvėje pasiskirstę visiškai skirtingai, tačiau metodai rekomenduoja panašų klasterių skaičių.