VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFOMATIKOS FAKULTETAS

DUOMENŲ MOKSLO BAKALAURAS

DIMENSIJOS MAŽINIMAS KLASTERIZAVIME

Laboratorinis darbas

Atliko: Simona Gelžinytė,

Ugnė Kniukštaitė, Rugilė Bagdonaitė

duomenų mokslas 3 k.

Vilnius, 2023

TURINYS

ĮVADAS

Tikslas

Išskirti tiriamoje duomenų aibėje klasterius bei apibrėžti susidariusių klasterių specifiką.

Uždaviniai

1. Pasirinkti ir įvardinti, pagal kokius požymius bus atliekamas klasterizavimas.
2. Naudojant empirinį, Elbow ir vidutinio silueto metodą įvertinti optimalų klasterių skaičių.
3. Suklasterizuoti duomenis naudojant k - means, hierarchinį ar DBSCAN klasterizavimo algoritmus visiems duomenis aprašantiems požymiams.
4. Suklasterizuoti ir vizualizuoti duomenis, gautus panaudojus dimensijos mažinimo algoritmus bei pagal kitus pasirinktus požymius.
5. Pateikti susidariusių klasterių aprašomąsias statistikas ir palyginti, kas pasikeitė klasterizavus originalius duomenis ir sumažinus dimensiją bei pagal pasirinktus požymius.

Duomenys

Darbe naudoti normuoti duomenys pagal min ir max reikšmes iš antro laboratorinio darbo - duomenų rinkinys apie diabetą. Iš viso yra 768 stebėjimai. Ankstesniame darbe buvo pastebėta, kad duomenų aibėje nėra praleistų reikšmių, tačiau kai kurie įrašai neatitinka logiškos kintamųjų skalės (pvz. kraujo spaudimas ar KMI lygūs 0), todėl darėme prielaidą, kad praleistos reikšmės buvo užpildytos 0. Atsižvelgiant į tai ir siekiant gauti tikslesnius rezultatus, įrašus, kuriuose KMI ir kraujospūdis buvo lygūs 0, pašalinome. Iš viso tokių netinkamų stebėjimų buvo 39. Dėl kitų įrašų buvo padaryta prielaida, jog yra teisingi. Buvo rasta 111 sąlyginių ir 16 tikrųjų išskirčių.

Šiame darbe pritaikėme 3 klasterizavimo metodus 3 duomenų rinkiniams – visiems duomenims, reikšmingoms kovariantėms bei sumažinus dimensiją iki 2.

Reikšmingos kovariantės buvo atrinktos pritaikius logistinį modelį nustatyti, ar žmogus serga, ar ne ir buvo gautas toks rezultatas:

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.8170408  0.8117630 -10.862  < 2e-16 ***
Pregnancies               0.1150347  0.0330853   3.477 0.000507 ***
Glucose                   0.0336410  0.0037572   8.954  < 2e-16 ***
BloodPressure            -0.0100157  0.0086060  -1.164 0.244501
SkinThickness             0.0007060  0.0070117   0.101 0.919797
Insulin                  -0.0010220  0.0009144  -1.118 0.263709
BMI                       0.0965139  0.0167561   5.760 8.42e-09 ***
DiabetesPedigreeFunction  0.9995318  0.3061460   3.265 0.001095 **
Age                       0.0179745  0.0097612   1.841 0.065558 .
```

1 pav. Reikšmingos kovariantės gautos pritaikius logistinį modelį.

Gautos reikšmingos kovariantės: nėštumų skaičius, gliukozės kiekis, KMI ir diabeto kilmės funkcija.

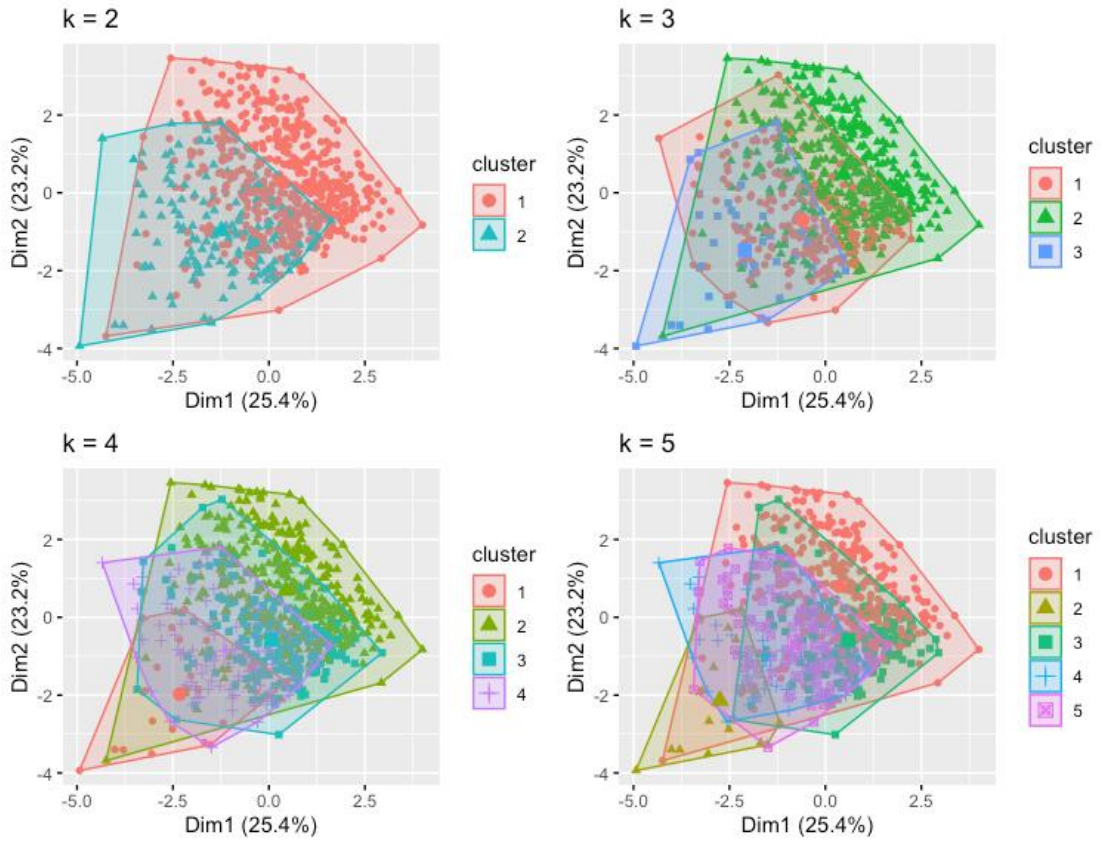
DIMENSIJOS MAŽINIMAS KLASTERIZAVIME

k – vidurkių klasterizavimo metodas

k - vidurkių metodas kiekvienam klasteriui apskaičiuoja centroidą (svorio centrą) ir kiekvieną objektą priskiria artimiausiam centroidui. Ši procedūra kartojama iki tol, kol tenkinamas sustojimo kriterijus. Centroidų (vadinasi ir klasterių) kiekis turi būti pasirenkamas iš anksto. Centroidai įprastai inicializuojami atsitiktinai. Šis klasterizavimo metodas lengvai interpretuojamas, gali būti taikomas dideliems duomenų kiekiams. [6]

Visam duomenų rinkiniui

Pradžioje klasterizavimą pritaikėme visam duomenų rinkiniui. Duomenys buvo normuoti pagal min - max metodą. Klasterių skaičius buvo pasirinktas 2,3,4 ir 5. Kiekvieną kartą algoritmas kūrė 25 konfigūracijas ir atrinko po vieną geriausią iš jų. Jei yra daugiau nei dvi dimensijos (kintamieji), braižant grafiką yra atliekama pagrindinių komponenčių analizė (PCA), t. y. taškai nubraižomi pagal dvi pirmąsias pagrindines komponentas, kurios paaiškina didžiąją dispersijos dalį.



2 pav. Klasterizavimo rezultatai, naudojant visus požymius

1 lentelė. Informacija apie gautus klasterius

	k = 2	k = 3	k = 4	k = 5
Klasterių dydžiai	78%, 22%	4%, 32%, 64%	52%, 15%, 29%, 4%	2%, 19%, , 25%, 54%
R^2	56,3 %	76%	82,7%	86,4%

Matome, jog pasirinkus klasterių skaičių 2, jie stipriai persidengia ir tikslumas šiek tiek didesnis nei 56%, didinant klasterių skaičių, didėja ir tikslumas.

2 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	401	167
2	77	84

Matome, jog 84 proc. nesergančių žmonių pateko į 1 klasterį, tačiau šiame klasteryje taip pat daug (67 proc.) sergančių pacientų.

3 lentelė. Klasterių aprašomoji statistika, kai klasterių skaičius = 2

		Vidurkis	Mediana	Min	Max
1	Nėštumas	4	3	0	17
	Gliukozė	115,25	111	0	199
	Kraujo spaudimas	72	72	24	122
	Odos storis	18,76	20	0	99
	Insulinas	35	0	0	145
	KMI	32	31,20	18,20	67,10
	Diabeto f-ja	0,44	0,34	0,08	2,42
	Amžius	33	29	21	81
2	Nėštumas	4	3	0	14
	Gliukozė	141,48	138	91	198
	Kraujo spaudimas	73	72	40	110
	Odos storis	31,18	32	7	51
	Insulinas	256	207	144	846
	KMI	35	34,5	19,60	57,30
	Diabeto f-ja	0,6	0,53	0,12	2,33
	Amžius	34	30	21	63

Empirinis skaičiavimas

Optimalus klasterių skaičius paskačiuojamas pagal formulę:

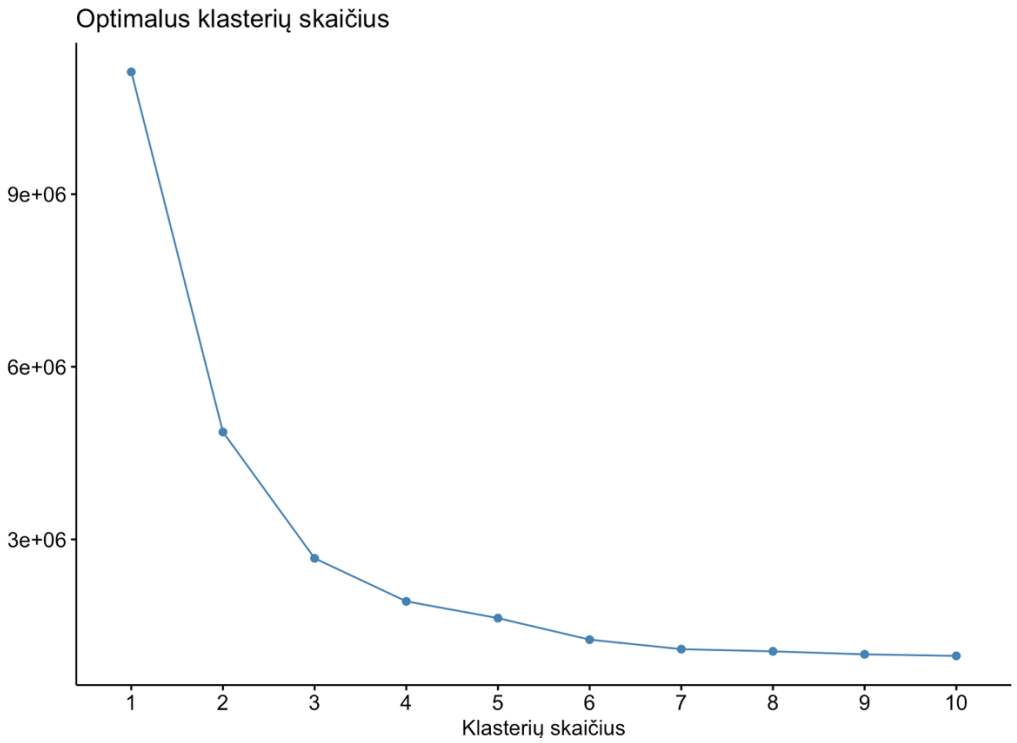
$$N_{klast} \approx \sqrt{n/2}$$

kur n – kovariančių skaičius

Pritaikę formulę gauname: $\sqrt{8/2} = 2$, vadinasi optimalus klasterių skaičius – 2.

Alkūnės (elbow) metodas

Optimaliam klasterių skaičiui rasti alkūnės metodu visiems klasteriams suskaičiuojamos ir į bendrą sumą sudedamos Euklidinių atstumų nuo klasterio vidurkio taško kvadratų sumos. Siekiant surasti optimalų klasterių skaičių nubraižoma šio gauto dydžio kreivė pagal klasterių skaičių k ir ieškoma linkio taško.



3 pav. Klasterių skaičiaus paieška alkūnės metodu

Kadangi linkio vieta grafike paprastai laikoma tinkamo klasterio skaičiaus rodikliu, iš grafiko matome, jog minimalus klasterių skaičius yra 3.

4 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	16	22
2	136	97
3	326	132

Į 3 klasterį patenka apie 68 proc. visų sveikų pacientų, tačiau pirmąjį klasterį sudaro tik 5 proc. visų stebinių.

5 3 lentelė Klasterių aprašomoji statistika, kai klasterių skaičius 3

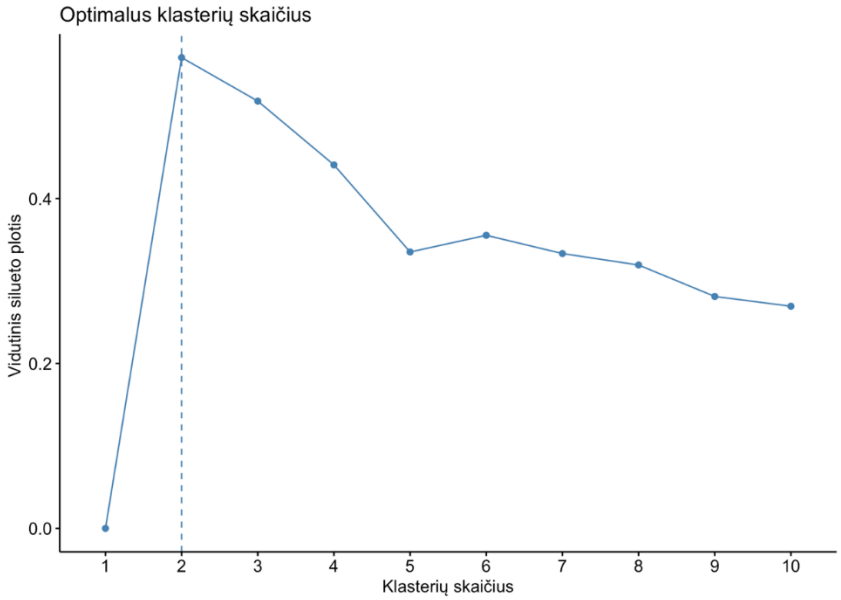
		Vidurkis	Mediana	Min	Max
1	Nėštumas	4	3	0	12
	Gliukozė	158,45	156	105	197
	Kraujo spaudimas	72	70	46	90
	Odos storis	32,26	33	7	49
	Insulinas	441,29	408,5	300	846
	KMI	35,11	35,05	19,6	52,3
	Diabeto f-ja	0,57	0,5	0,13	2,33
	Amžius	35	30	21	60
2	Nėštumas	4	2	0	17
	Gliukozė	129,5	125	82	198
	Kraujo spaudimas	71,5	72	30	110
	Odos storis	30,38	30	7	60
	Insulinas	159,7	150	88	293
	KMI	34,14	34	21,20	67,1
	Diabeto f-ja	0,53	0,46	0,09	2,29
	Amžius	32	28	21	63
3	Nėštumas	4	3	0	14
	Gliukozė	113,64	108	0	199
	Kraujo spaudimas	73	72	24	122
	Odos storis	16,09	16,5	0	99
	Insulinas	15,76	0	0	90
	KMI	31,4	30,95	18,20	59,4
	Diabeto f-ja	0,44	0,33	0,08	2,42
	Amžius	34	30	21	81

Matome, jog 1 klasteryje stebėjimų medianų reikšmės didžiausios iš visų grupių gliukozės, odos storio, insulino, KMI diabeto atsiradimo funkcijos kovariantėms. Daugiausiai mažiausios medianos stebėjimų yra 3 klasteryje, kur mažiausia mediana pastebima gliukozės, odos storio, insulino, KMI, diabeto atsiradimo funkcijos kovariantėms.

Vidutinio silueto metodas

Silueto koeficientas lygina kiekvieno objekto panašumą su savo paties klasterio objektais lyginant su panašiausiu kito klasterio objektu. Optimalaus klasterių skaičiaus radimo metode, paremtame silueto koeficientais, kiekvienam klasterių skaičiui k skaičiuojamas vidutinis silueto

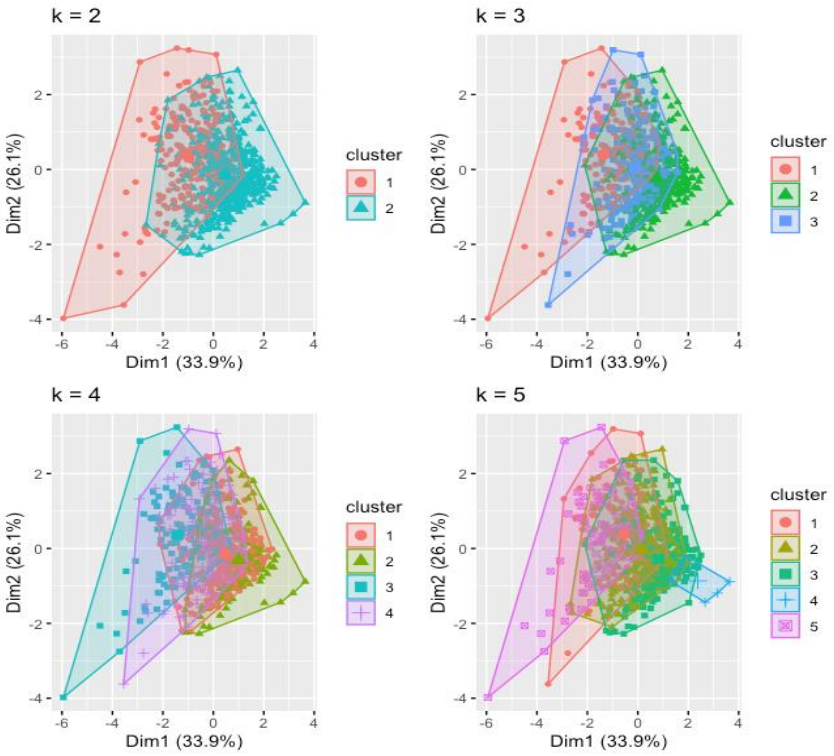
koeficientas ir kaip optimalus klasterių skaičius pasirenkamas toks k, su kurio gaunama didžiausia šio dydžio reikšmė.



4 pav. Klasterio skaičiaus paieška vidutinio silueto metodu

Didžiausia reikšmė grafike paprastai laikoma tinkamo klasterio skaičiaus rodikliu, mūsų atveju, tai yra 2, kitas optimalus klasterių skaičius yra 3.

Reikšmingoms kovariantėms



5 pav. Klasterizavimo rezultatai, kai naudojamos reikšmingos kovariantės

6 lentelė. Informacija apie klasterius

	k= 2	k = 3	k = 4	k = 5
Klasterių dydžiai	33%, 67%	19%, 41%, 40%	36%, 17%, 16%, 31%	20%, 36%, 30%, 1%, 13%
R^2	61,1%	76,4%	82,6%	87%

Matome, jog pasirinkus klasterių skaičių 2, jie persidengia ir tikslumas 61,1%, didinant klasterių skaičių, didėja ir tikslumas, tačiau persidengimas lieka.

7 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	90	149
2	388	102

Iš lentelės matome, jog 51 proc. visų sergančių pacientų pateko 1 klasterį, o 81 proc. sveikų pacientų į 2.

8 lentelė. Klasterių aprašomoji statistika, kai klasterių skaičius 2

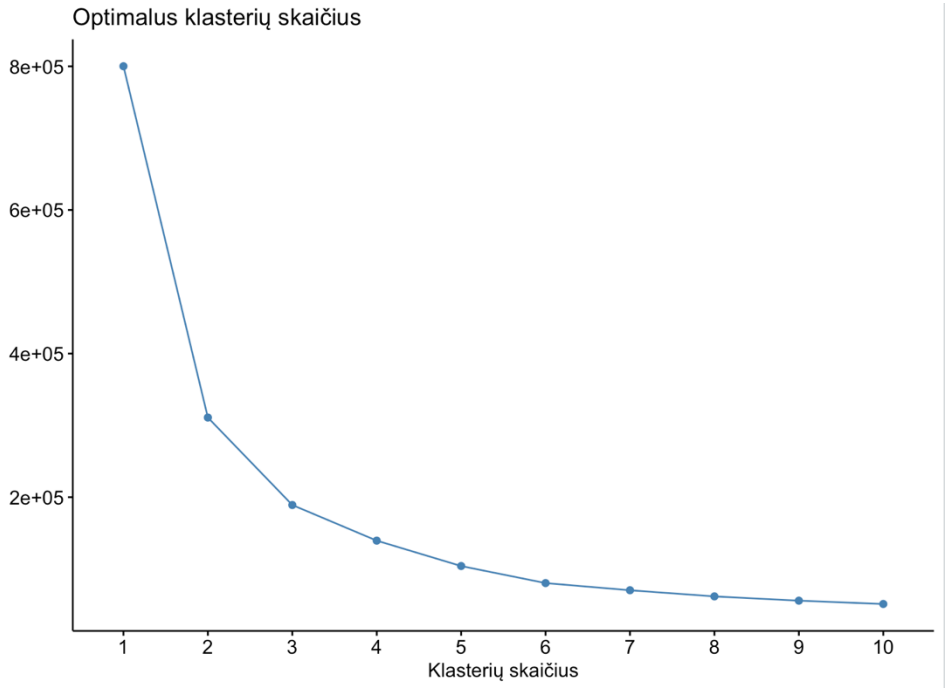
		Vidurkis	Mediana	Min	Max
1	Nėštumas	5	4	0	17
	Gliukozė	158,08	154	129	199
	KMI	34,5	33,7	21	67,10
	Diabeto f-ja	0,52	0,4	0,09	2,42
2	Nėštumas	4	2	0	14
	Gliukozė	102,98	105	0	131
	KMI	31,48	31,2	18,2	57,3
	Diabeto f-ja	0,45	0,37	0,08	1,7

Matome, jog pirmajame klasteryje susigrupavo stebėjimai, turintys didžiausią medianą visoms kovariantėms.

Empirinis skaičiavimas

Pritaikę formulę gauname: $\sqrt{4/2} \approx 1$, vadinasi optimalus klasterių skaičius – 1.

Alkūnės (elbow) metodas



6 pav. Klasterių skaičiaus paieška alkūnės metodu

Iš grafiko matome, jog minimalus klasterių skaičius yra 3.

9 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	37	103
2	256	43
3	185	105

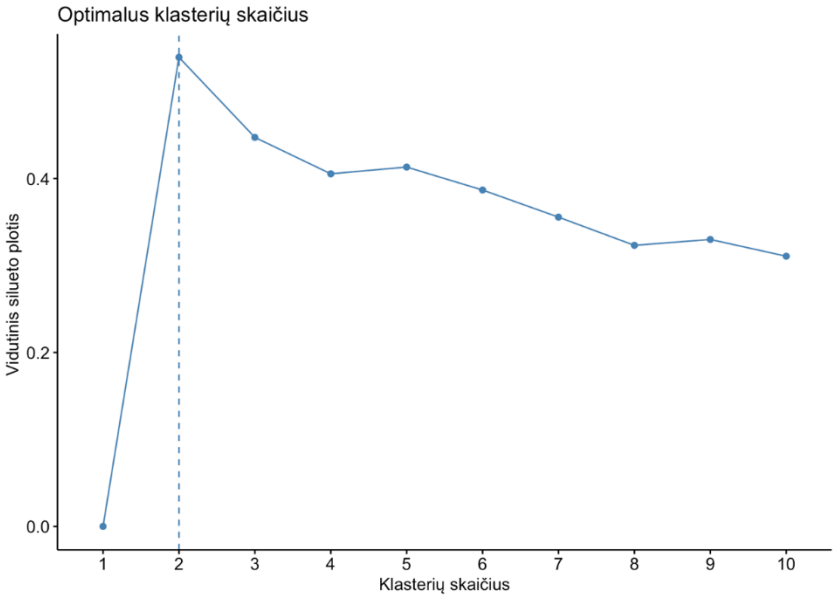
Iš lentelės matome, jog 2 klasterį 87 proc. sudaro sveiki pacientai.

10 lentelė. Klasterių aprašomoji statistika, kai klasterių skaičius 3

		Vidurkis	Medina	Min	Max
1	Nėštumas	4	3	0	15
	Gliukozė	126,86	125	109	149
	KMI	32,78	32,4	19,6	67,1
	Diabeto f-ja	0,46	0,37	0,09	2,29
2	Nėštumas	3	2	0	14
	Gliukozė	92,06	95	0	109
	KMI	30,90	30,8	18,2	55
	Diabeto f-ja	0,45	0,37	0,08	1,7
3	Nėštumas	5	5	0	17
	Gliukozė	170,91	168,5	150	199
	KMI	35,18	34,4	21	59,4
	Diabeto f-ja	0,56	0,43	0,09	2,42

Iš lentelės galime matyti, jog 3 klasterio medianos reikšmė didesnė už kitų klasterių visom naudojamoms kovariantėms, o 2 - mažiausia, 1 tarpinės medianos reikšmės.

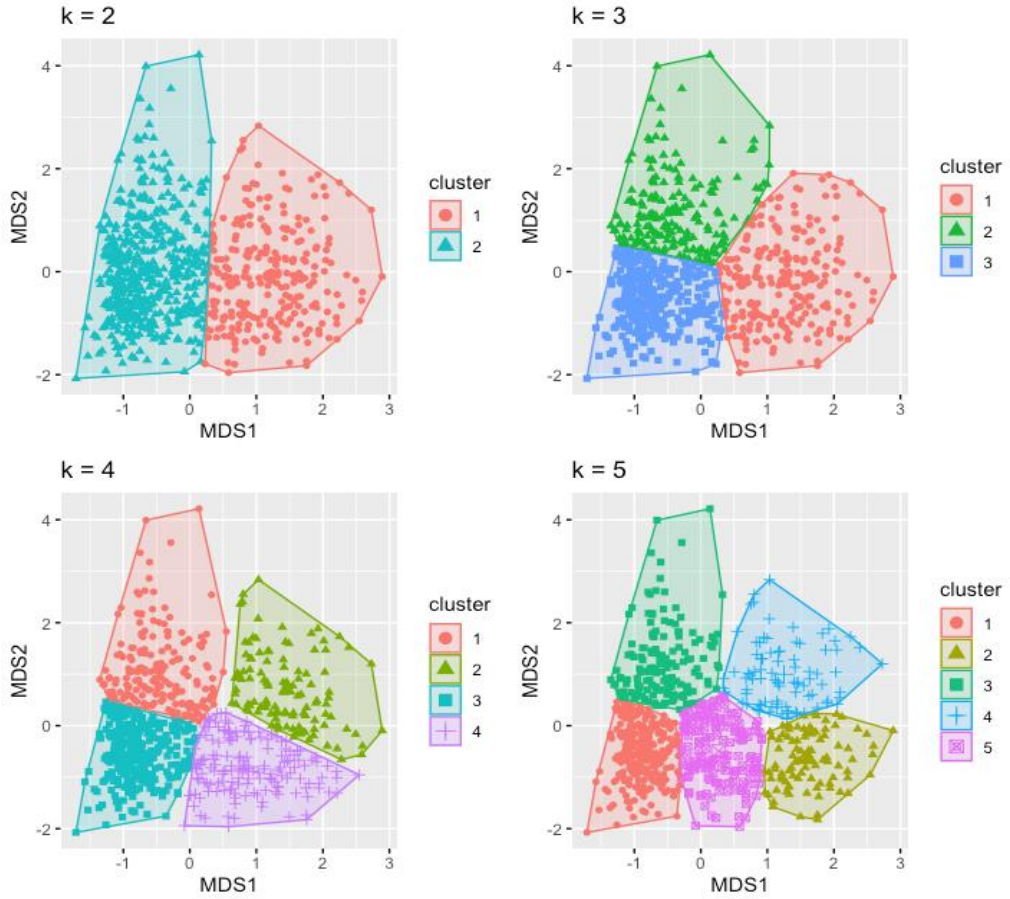
Vidutinio silueto metodas



7 pav. Klasterių skaičiaus paieška vidutinio silueto metodu

Mūsų atveju, minimalus klasterių skaičius yra 2, kitas optimalus klasterių skaičius yra 3.

Sumažintos dimensijos



8 pav. Klasterizavimo rezultatai, naudojant sumažintos dimensijos duomenis

11 lentelė. Informacija apie klasterius

	k= 2	k = 3	k = 4	k = 5
Klasterių dydžiai	35%, 65%	32%, 43%, 25%	25%, 38%, 15%, 22%	33%, 13%, 12%, 22%, 20%
R^2	45,6%	64%	71,6%	77,4%

Naudojant MDS dimensijos mažinimo algoritmą, matome, jog klasteriai gerai atsiskiria (nepersidengia), tačiau R^2, kuris parodo kaip tiksliai atsiskiria klasteriai yra nedidelis – 45,6%. Didėjant klasterių skaičiui, didėja ir tikslumas.

12 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	120	133
2	358	118

Iš lentelės matome, jog 75 proc. sveikų pacientų grupavosi 2 klasteryje ir apie 53 proc. sergančių 1.

13 lentelė. Klasterių aprašomoji statistika, kai klasterių skaičius yra 2

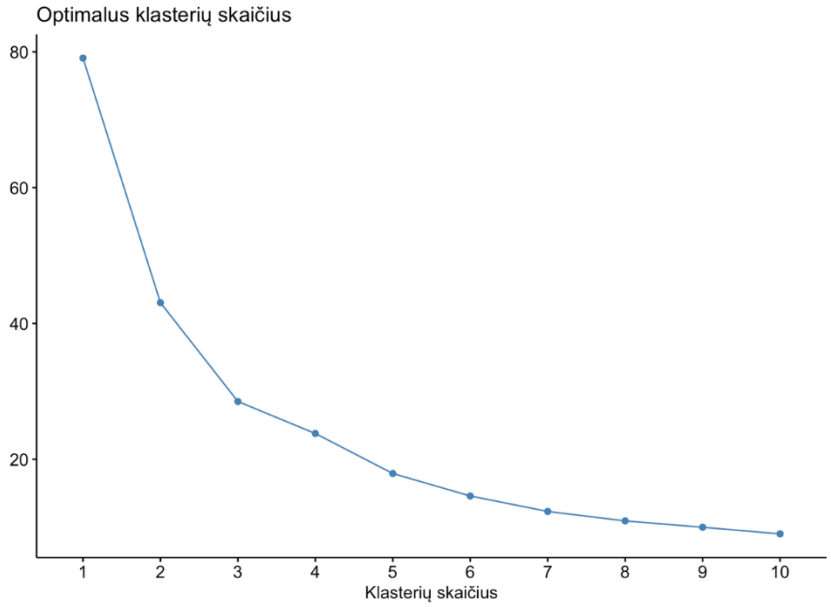
		Vidurkis	Medina	Min	Max
1	MDS1	0,3	0,29	0,3	0,74
	MDS2	-0,02	-0,03	-0,3	0,59
2	MDS1	-0,16	-0,18	-0,44	0,08
	MDS2	0,01	-0,01	-0,43	0,88

Iš lentelės matome, jog nėra vienareikšmio klasterių atsiskyrimo pagal medianos reikšmes.

Empirinis skaičiavimas

Pritaikę formulę gauname: $\sqrt{2/2} = 1$, vadinasi optimalus klasterių skaičius – 1.

Alkūnės (elbow) metodas



9 pav. Optimalaus klasterių skaičiaus radimas alkūnės metodu

Iš grafiko matome, jog minimalus klasterių skaičius yra 3.

14 lentelė. Klasterizavimo klasių apsiskirstymas

Klasterio nr.	Klasių klasterizavimas	
	0	1
1	37	103
2	256	43
3	185	105

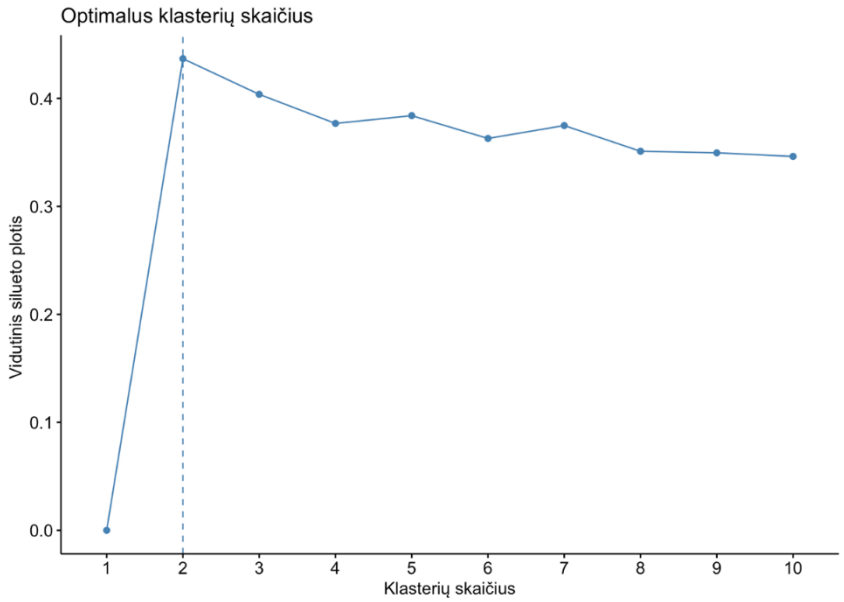
Matome, jog 86 proc. visų stebinių 2 klasterio yra sveiki, o 74 proc. 1 klasterio stebinių yra sergantys.

15 lentelė. Klasterių aprašomoji statistika, kai klasterių skaičius 3

		Vidurkis	Mediana	Min	Max
1	MDS1	0,02	-0,05	-0,32	0,74
	MDS2	0,01	0	-0,32	0,54
2	MDS1	-0,09	-0,19	-0,44	0,57
	MDS2	-0,1	-0,12	-0,43	0,42
3	MDS1	0,16	0,18	-0,27	0,7
	MDS2	0,19	0,19	-0,23	0,88

Iš lentelės galime pastebėti, jog 1 grupėje grupavosi tarpinės medianos pacientai, 2 su mažiausiomis reikšmėmis, o 3 grupavosi didžiausių reikšmių pacientai.

Vidutinio silueto metodas



10 pav. Optimalaus klasterių skaičiaus radimas vidutinio silueto metodu

Minimalus klasterių skaičius yra 2.

Hierarchinis klasterizavimo metodas

Hierarchinis klasterizavimo metodas yra vienas iš klasterizavimo algoritmų, kuris grupuoja objektus į klasterius pagal jų panašumus. Šis metodas suskirsto objektus į hierarchines struktūras, kurias galima atvaizduoti kaip dendrogramą.

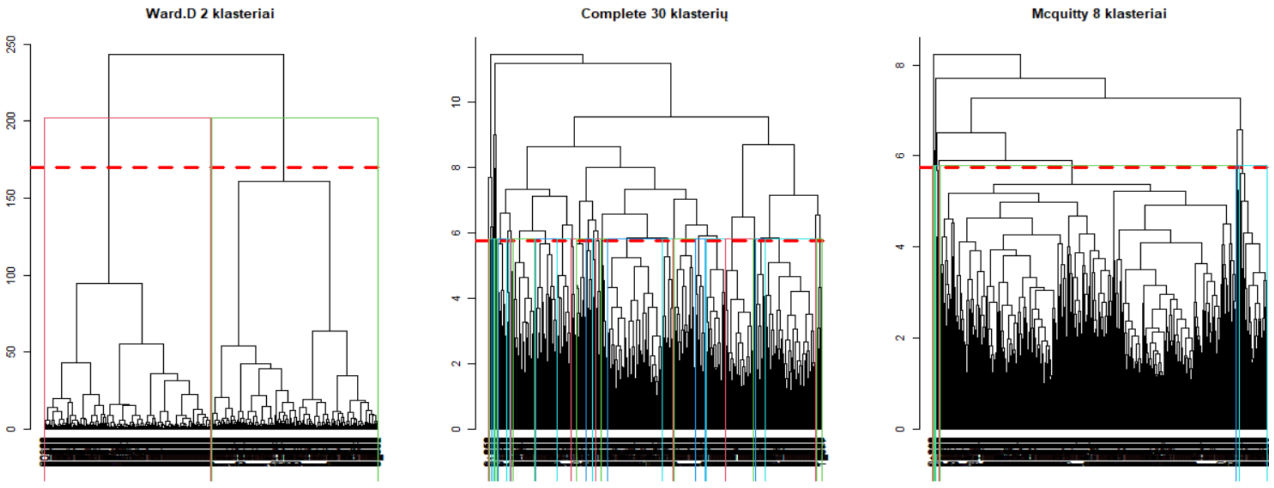
Šis metodas pradeda nuo to, kad kiekvienas objektas yra laikomas atskiru klasteriu. Tada panašumo matas apskaičiuojamas tarp visų galimų porų klasterių, naudojant atitinkamą metriką. Įvertinimo metrikai dažniausiai naudojami atstumai tarp taškų, pvz., Euklido atstumas arba kosinusinis atstumas.

Tada šie klasteriai yra sujungiami, kad būtų gauti nauji klasteriai. Šis procesas tęsiasi, kol visi objektai yra suskirstyti į vieną bendrą klasterį. Ši procedūra vadinama aglomeraciniu klasterizavimu, nes klasteriai yra susijungti nuo apačios į viršų. Alternatyvus metodas yra dalijantis klasterizavimas, kai pradžioje visi objektai yra vieno klasterio, tada jie dalinami į mažesnius klasterius, kol yra pasiektas reikiamas skaičius klasterių. Tolimesniame tyrime bus taikomas pastarasis metodas.

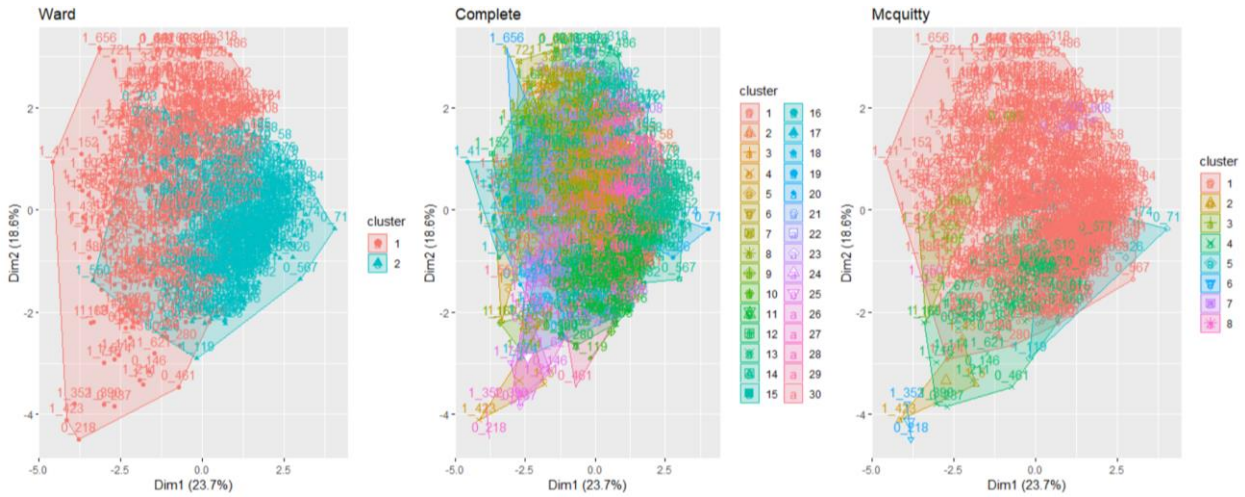
Dendrogramoje kiekvienas aukštesnis lygis reprezentuoja klasterių grupes, o žemiausias lygis yra pradiniai objektai. Atstumas dendrogramoje tarp dviejų taškų atspindi atstumą tarp jų klasterių pagal turimą metriką. [5]

Visam duomenų rinkiniui

Iš pradžių klasterizavimas buvo pritaikytas visam duomenų rinkiniui, normuotiems duomenims. Buvo taikomi 3 atstumų skaičiavimo metodai – Euklidinis, Manheteno, maksimumo ir 3 klasterizavimo metodai – Ward.D, Complete, Mcquitty. Su Euklidiniu atstumo skaičiavimo metodu Ward.D dendograma parodė, kad optimaliausias klasterių skaičius 2, Complete – 30, o Mcquitty – 8 klasteriai.



11 pav. Dendrogramos normuotiems duomenims skaičiuojant Euklidinius atstumus.



12 pav. Klasterizavimas normuotiems duomenims su Euklidiniu atstumų skaičiavimu.

16 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

71 lentelė. Klasterizuotų klasių pasiskirstymas su Euklidiniu atstumu ir Ward.D algoritmu

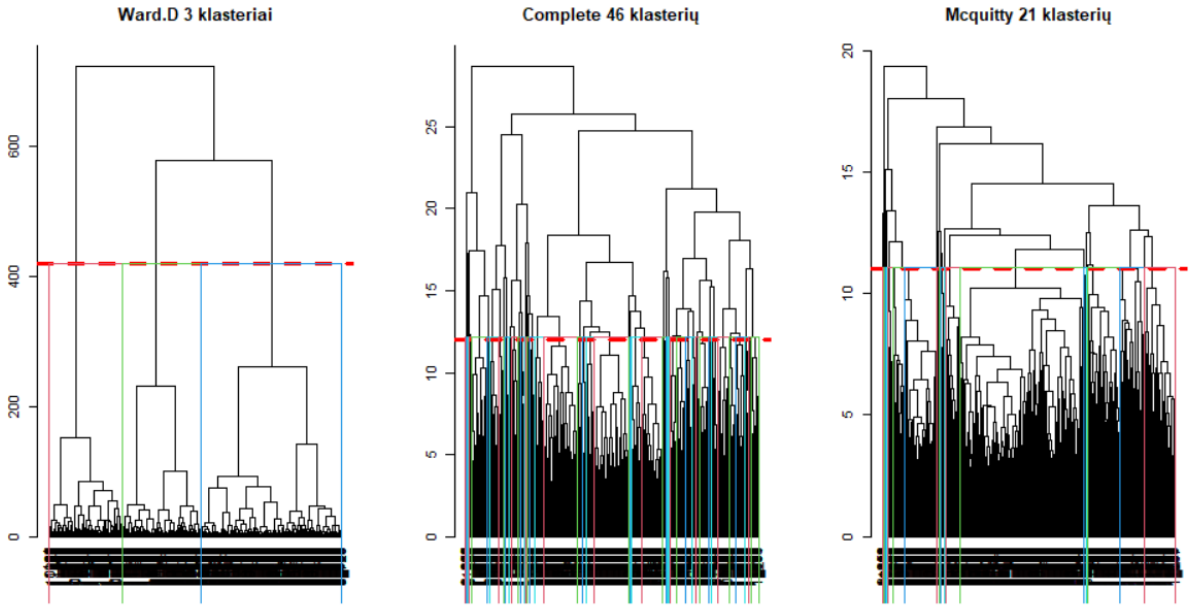
Klasterio nr.	Ward klasių klasterizavimas	
	0	1
1	211	187
2	267	64

18 lentelė. Klasterizuotų klasių pasiskirstymas su Euklidiniu atstumu ir Mcquitty algoritmu

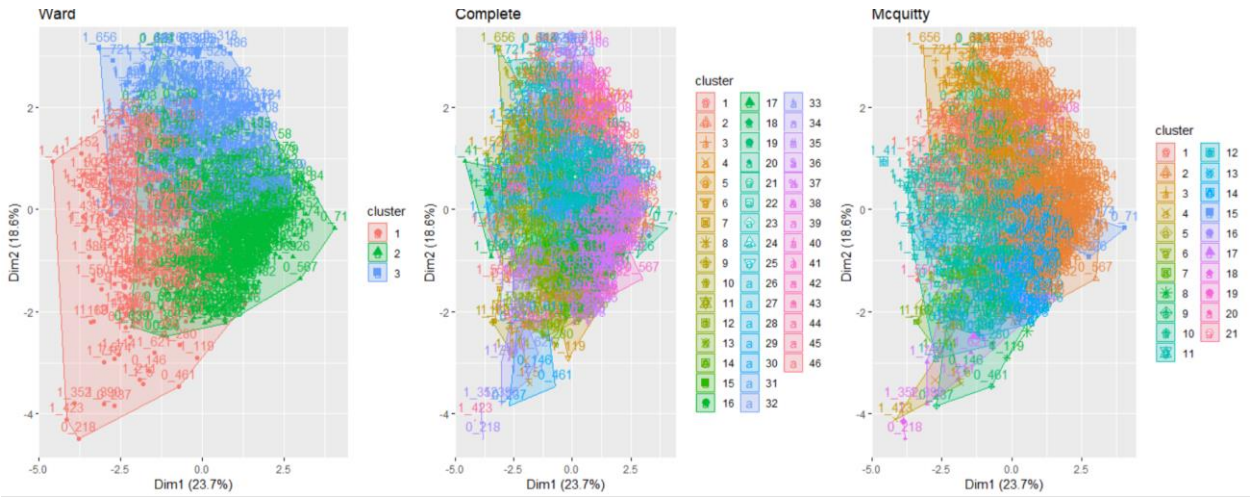
Klasterio nr.	Mcquitty klasių klasterizavimas	
	0	1
1	113	75
2	190	8
3	95	71
4	64	50
5	16	47

Galima pastebėti, kad klasteriai persidengia, Complete ir Mcquitty metodai didžiają dalį stebėjimų sudėjo į vieną klasterį, o Ward metodas pasirodė geriausiai – 2 klasteryje daugiausiai susigrupavo nesergančių, o 1 klasteryje daugiausiai sergančių pacientų pakliuvo, t. y. 1 klasteryje pakliuvo apie 75 proc. visų sergančių pacientų, o 2 apie 56 proc. visų sveikų asmenų.

Toliau tas pats pritaikyta skaičiuojant atstumą Manheteno metodu. Dendrogramos parodė, kad optimalus klasterių skaičius Ward.D algoritmu yra 3, Complete – 46, o Mcquitty 21 klasteris.



13 pav. Dendrogramos normuotiems duomenims skaičiuojant Manheteno atstumus.



14 pav. Klasterizavimas normuotiems duomenims su Manheteno atstumų skaičiavimu.

Klasteriai stipriai persidengia su Complete ir Mcquitty metodais, šiek tiek geriau klasteriai atsiskiria taikant Ward.D metodą.

19 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

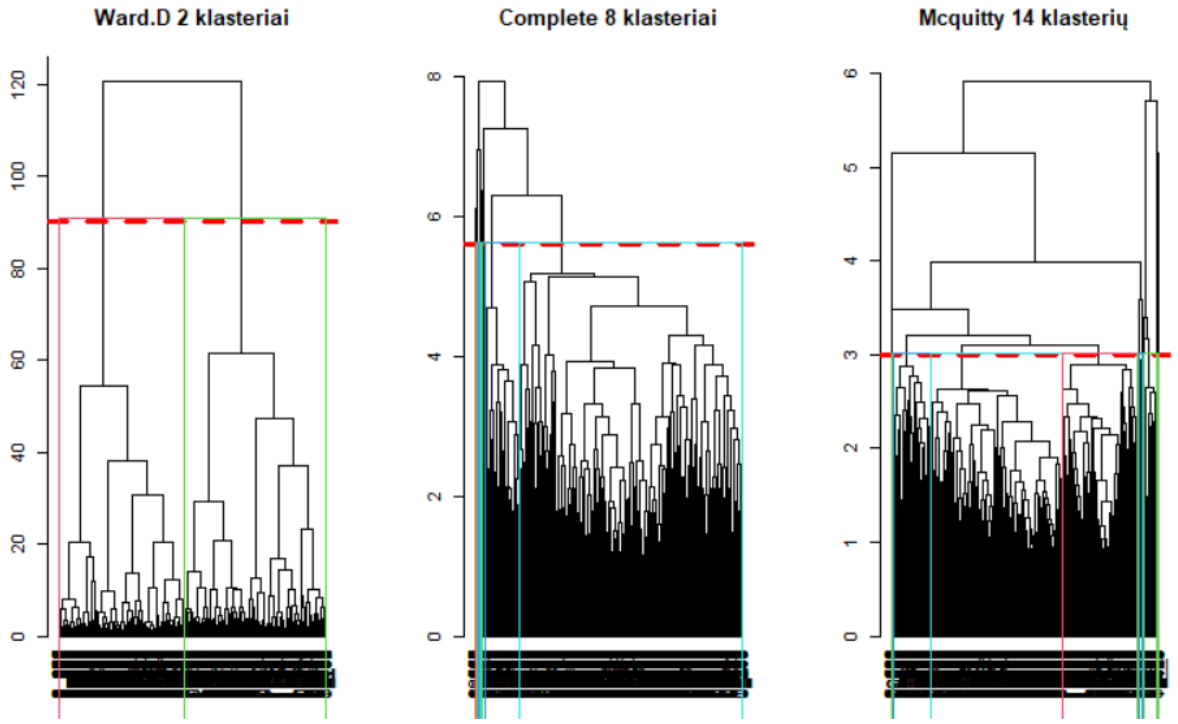
20 lentelė. Klasterizuotų klasių pasiskirstymas su Manheteno atstumu ir Ward.D metodu

Klasterio nr.	Ward klasių klasterizavimas	
	0	1
1	13	170
2	352	0
3	113	81

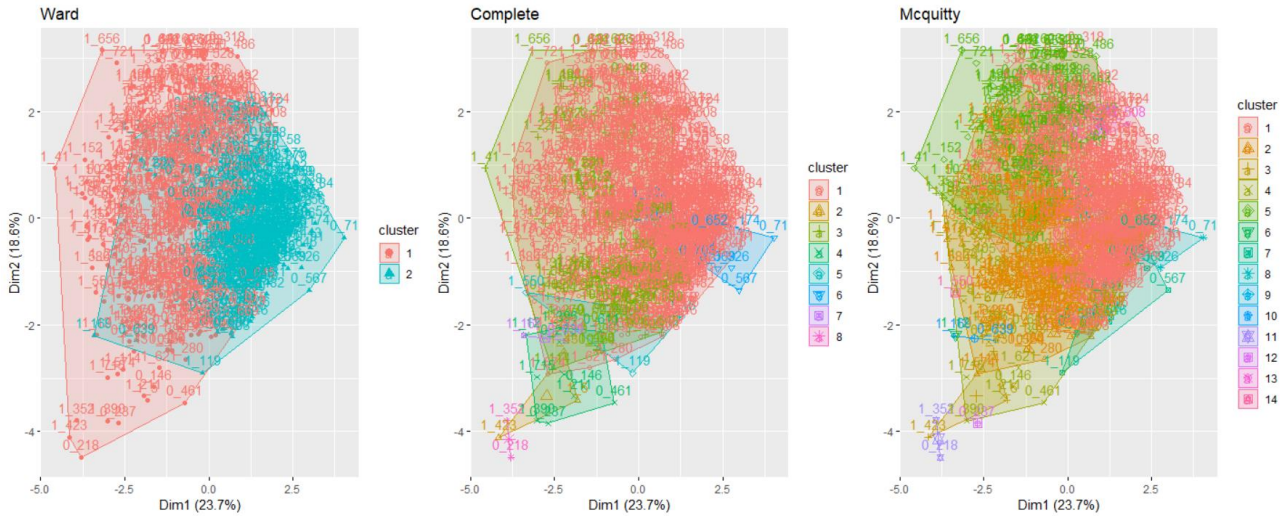
Šiuo atveju matome, kad geriausiai klasterizuoja Ward metodas, kurio 1 klasteryje didžioji dalis stebėjimų patenka sveiki asmenys, 2 klasteryje sergantys, o 3 yra ir sergančių, ir sveikų, t. y. į 1 klasterį patenką apie 68 proc. visų sveikų asmenų, o į 2 klasterį apie 74 proc. sergančių.

Complete ir Mcquitty metodų klasių pasiskirstymo nepateiksime, nes šie algoritmai išskiria labai daugiau klasterių.

Pakartojus bandymą atstumą skaičiuojant maksimumo metodu, gautas optimalus klasterių skaičius – 2, naudojant Ward.D metodą, 8 klasteriai su Complete metodu ir 14 su Mcquitty algoritmu.



15 pav. Dendrogramos normuotiems duomenims skaičiuojant maksimumo atstumus.



16 pav. Klasterizavimas normuotiems duomenims su maksimumo atstumų skaičiavimu.

Matome, jog su Complete ir Mcquitty metodais klasteriai stipriai persidengia, nėra aiškaus atsiskyrimo, o su Ward.D metodu klasteriai aiškiau yra atskiriami.

21 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

22 lentelė. Klasterizuotų klasių pasiskirstymas su maksimumo atstumu ir Ward.D metodu

Klasterio nr.	Ward klasių klasterizavimas	
	0	1
1	142	245
2	336	6

Galime pastebėti, jog apie 98 proc. sergančių pacientų patenka į 1 klasterį, o apie 70 proc. sveikų į 2 klasterį, naudojant Ward.D metodą. Tai geriausias gautas rezultatas, kai nagrinėjami visi požymiai. Tikslesnių atsiskyrimų nepastebėjome nei su Complete, nei su Mcquitty metodais.

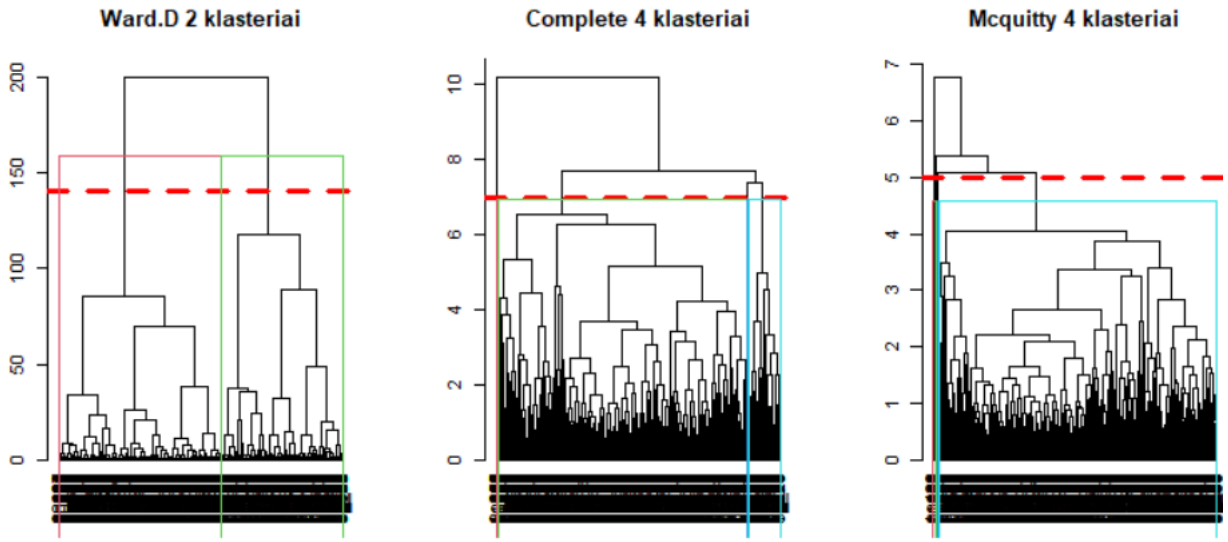
Toliau pateiktoje lentelėje galima matyti su maksimumo atstumais ir Ward metodu klasterizuotų normuotų duomenų skaitines charakteristikas pagal klasterius. Galime pastebėti, kad antro klasterio grupių medianos yra žemesnės nei pirmame klasteryje. Tai patvirtina, jog pirmajame klasteryje yra daugiau sergančių pacientų.

23 lentelė. Klasterio grupių aprašomoji statistika normuotiems duomenims.

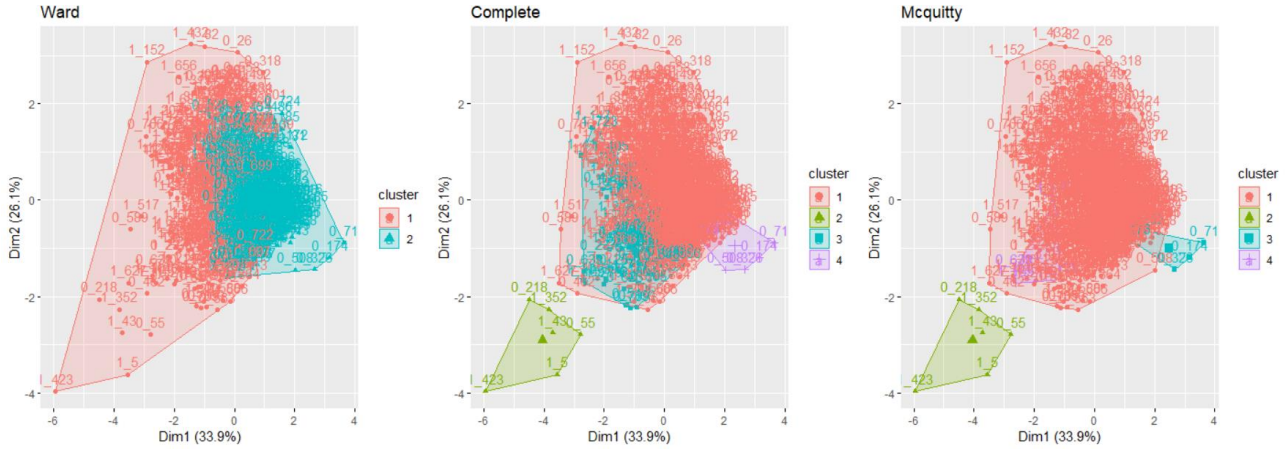
Klasterizavimas normuotiems duomenims						
Klasteris nr. 1						
	Vidurkis	Mediana	Min	Max	1Q	3Q
Nėštumų skaičius	0,40	0,34	-1,15	3,91	-0,55	1,23
Gliukozės koncentracija	0,44	0,37	-1,99	2,42	-0,31	1,15
Kraujo spaudimas	0,24	0,22	-2,62	3,36	-0,35	0,78
Odos storis	0,09	0,35	-1,37	4,93	-1,37	0,92
Insulinas	0,23	-0,23	-0,72	6,52	-0,72	0,82
KMI	0,23	0,16	-1,87	3,91	-0,37	0,78
Diabeto atsiradimo funkcija	0,20	-0,15	-1,17	5,87	-0,65	0,74
Amžius	0,49	0,40	-1,05	4,06	-0,37	1,16
Klasteris nr. 2						
Nėštumų skaičius	-0,45	-0,55	-1,15	2,13	-0,85	0,04
Gliukozės koncentracija	-0,50	-0,54	-3,75	2,17	-0,93	-0,06
Kraujo spaudimas	-0,27	-0,35	-3,9	4,01	-1,00	0,29
Odos storis	-0,10	0,03	-1,37	2,07	-0,67	0,54
Insulinas	-0,26	-0,34	-0,72	1,66	-0,72	0,09
KMI	-0,6	-0,40	-2,07	5,03	-1,03	0,35
Diabeto atsiradimo funkcija	-0,23	-0,41	-1,19	2,39	-0,73	0,15
Amžius	-0,56	-0,71	-1,05	2,27	-0,96	-0,37

Reikšmingoms kovariantėms

Toliau klasterizavimą pritaikėme pagal logistinį modelį gautoms reikšmingoms kovariantėms – nėštumų skaičiui, gliukozės kiekiui, KMI ir diabeto kilmės funkcijai. Skaičiuojant atstumą Euklidiniu metodu dendrograma Ward.D metodu parodė, jog optimalus klasterių skaičius yra 2, Complete – 4, o Mcquitty metodas parodė, jog optimalus klasterių skaičius yra 4.



17 pav. Dendrogramos reikšmingoms kovariantėms skaičiuojant Euklidinius atstumus.



18 pav. Klasterizavimas reikšmingoms kovariantėms su Euklidiniais atstumais.

24 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

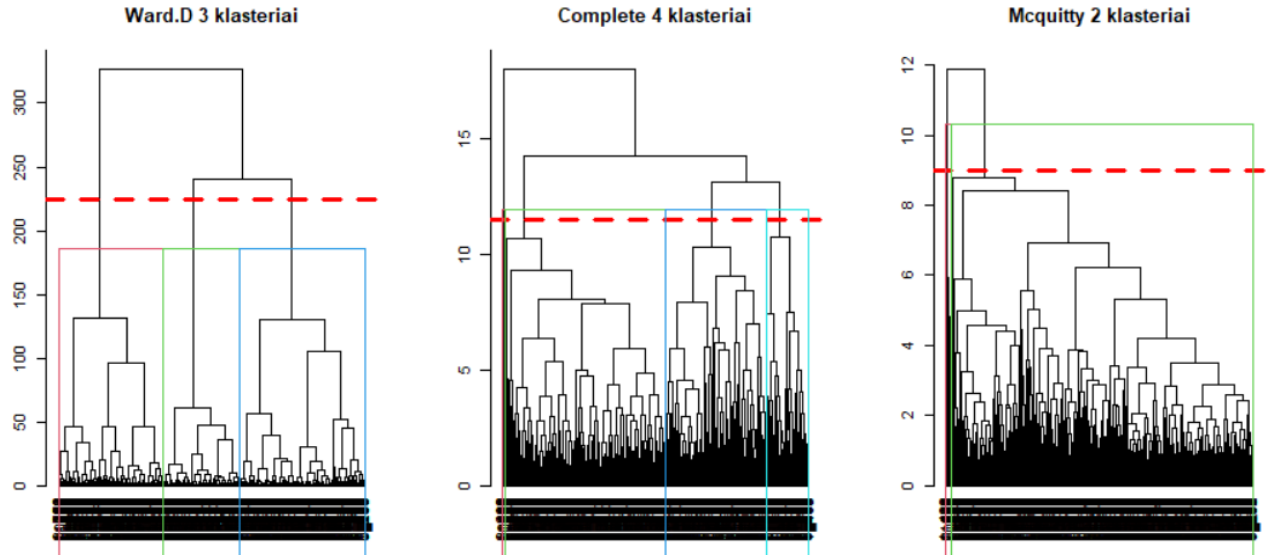
25 lentelė. Klasterizuotų klasių pasiskirstymas su Euklidiniais atstumais.

Klasterio nr.	Ward klasių klasterizavimas		Complete klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1	0	1
1	136	177	428	210	472	242
2	342	74	22	4	2	4
3			43	35	3	2
4			5	2	1	3

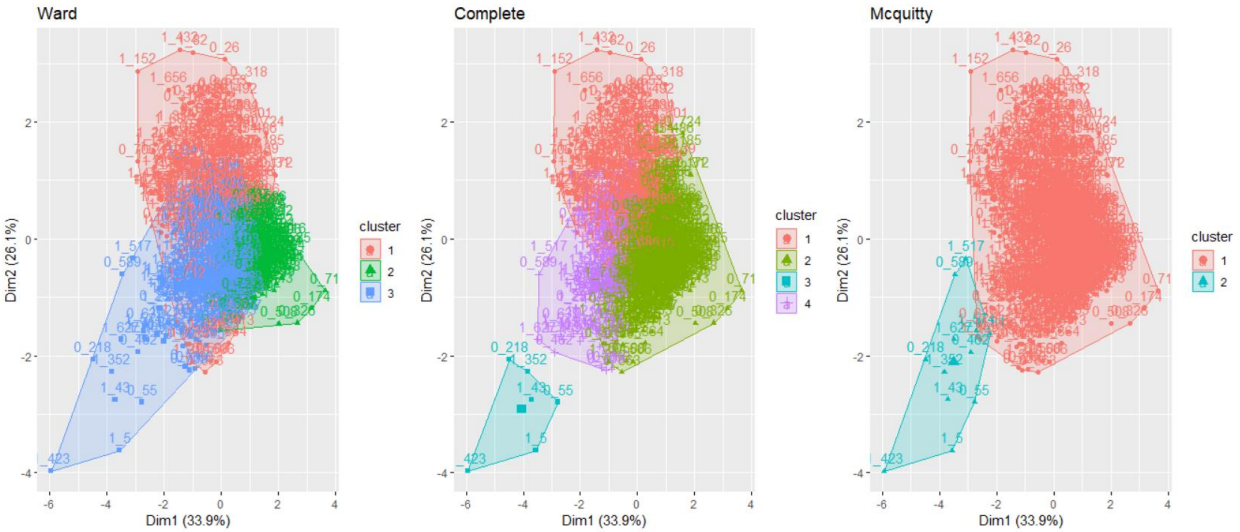
Atstumą skaičiuojant Euklidiniu metodu ir klasterizavimui pritaikius Ward metodą buvo gauta, kad 71,5 % nesergančių nukeliavo į 2 klasterį, o 70,5 % sergančių – į 1 klasterį, tačiau

klasteriai tarpusavyje smarkiai persidengia. Tuo tarpu Complete ir Mcquitty klasterizavimo rezultatai neparodė geresnių rezultatų – didžioji dalis pacientų pateko į 1 klasterį.

Toliau bandymas tęsiamas atstumą skaičiuojant Manheteno metodu. Buvo gauta, jog optimalus klasterių skaičius yra 3, 4, 2 atitinkamai gauti iš Ward.D, Complete ir Mcquitty metodų, sprendžiant iš dendogramų.



19 pav. Dendrogramos reikšmingoms kovariantėms su Manheteno atstumais.



20 pav. Klasterizavimas reikšmingoms kovariantėms su Manheteno atstumų skaičiavimu.

Galime pastebėti, jog geriausiai klasteriai atsiskiria su Mcquitty metodu, su kitais metodais pastebime persidengimą.

26 lentelė. Originalus klasių pasiskirstymas.

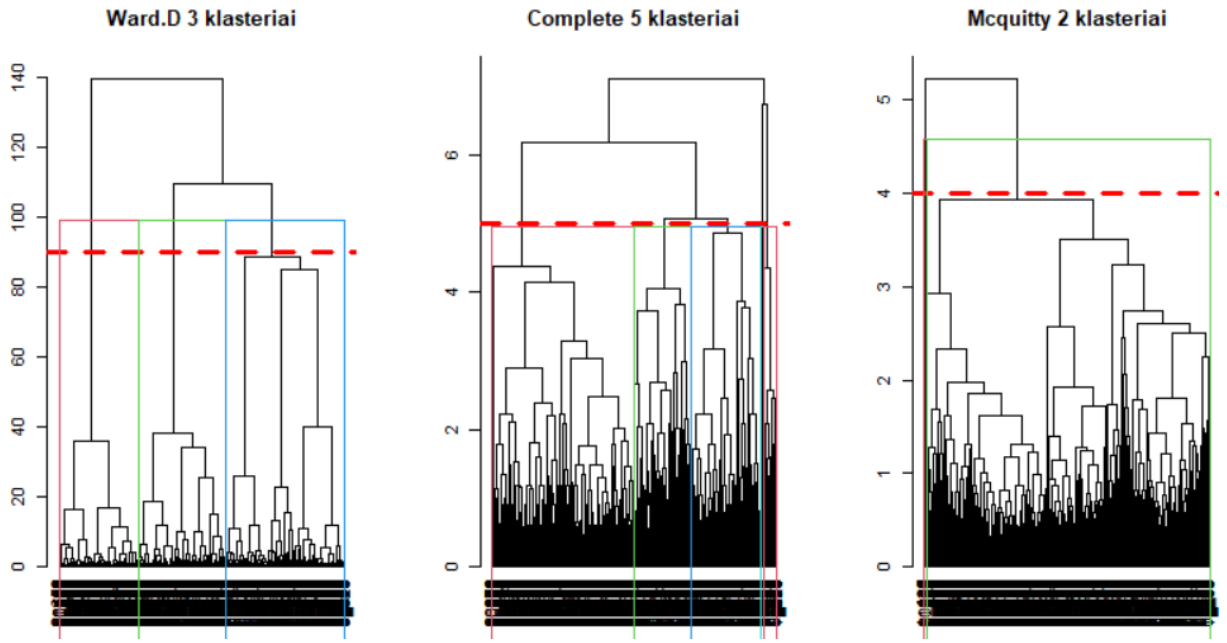
Originalus klasių pasiskirstymas	
0	1
478	251

27 lentelė. Klasterizuotų klasių pasiskirstymas su Manheteno atstumais.

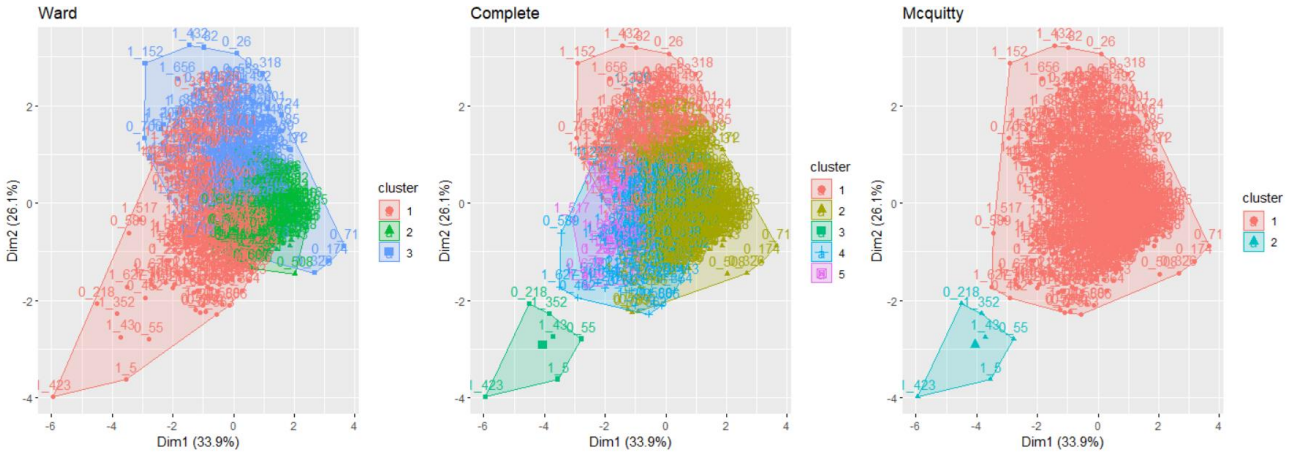
Klasterio nr.	Ward klasių klasterizavimas		Complete klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1	0	1
1	127	122	114	125	474	243
2	169	13	328	54	4	8
3	182	116	2	4		
4			34	68		

Iš grafikų ir duomenų pasiskirstymo lentelės galima pastebėti, kad klasės atskiriamos prasčiau naudojant Manheteno metodą atstumų skaičiavimui. Blogiausiai atsiskiria taikant Complete ir Mcquitty metodus, geresnis rezultatas pastebimas su Ward – didžiąją dalį (93 proc.) 2 klasterio sudaro sveiki pacientai, kitos grupės neturi tokio gero išskiriamumo.

Toliau pritaikius maksimumo metodą atstumų skaičiavimui, dendrogramos parodė, kad optimalus klasterių skaičius yra 2 naudojant Mcquitty metodą, 5 su Complete metodu ir 3 klasteriai pasitelkiant Ward.D metodą.



21 pav. Dendrogramos reikšmingoms kovariantėms su maksimumo atstumais.



22 pav. Klasterizavimas reikšmingoms kovariantėms su maksimumo atstumų skaičiavimu.

Taikant Ward.D ir Complete metodus grupės yra persidengusios, o Mcquitty grupes galima identifikuoti.

28 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

29 lentelė. Klasterizuotų klasių pasiskirstymas su maksimumo atstumais.

Klasterio nr.	Ward klasių klasterizavimas		Complete klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1	0	1
1	148	155	60	86	476	247
2	183	19	308	54	2	4
3	147	77	2	4		
4			98	81		
5			10	26		

Galime pastebėti, jog Mcquitty klasterizavimas nebuvo geras, nes apie 99 proc. stebinių pakliuvo į vieną klasterį, Complete metodo grupės neparodė jokio paciento atskiriamumo. Geriausiai iš šių trijų metodų pasirodė Ward.D metodas – 2 klasterį sudaro 91 proc. sveikų pacientų iš visų priskirtų asmenų 2 klasteriui.

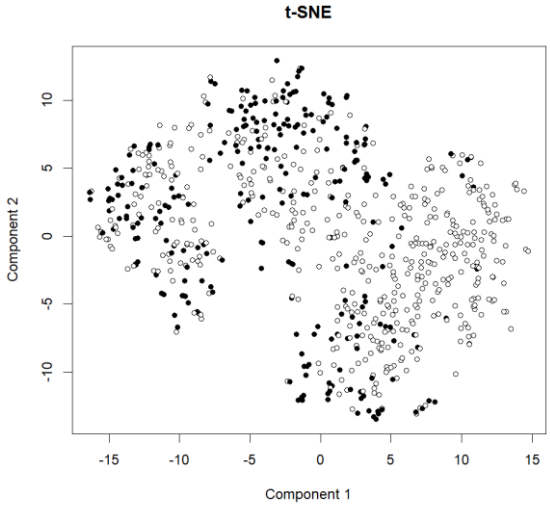
Iš gautų rezultatų matome, kad klasės atrinktos į klasterius dar prasčiau ir iš atliktų bandymų, galima pastebėti, kad geriausiai klases klasterizavo su reikšmingomis kovariantėmis Ward metodas su Euklidiniu atstumų skaičiavimo metodu, kurio klasterių aprašomosios statistikos pateiktos 20 lentelėje. Antrojo klasterio medianos vėl žemesnės nei pirmojo.

30 lentelė. Klasterio grupių aprašomoji statistika

Klasterizavimas reikšmingoms kovariantėms						
Klasteris nr. 1						
	Vidurkis	Mediana	Min	Max	1Q	3Q
Nėštumų skaičius	0,44	0,34	-1,15	3,91	-0,55	1,53
Gliukozės koncentracija	0,50	0,46	-1,99	2,42	-0,37	1,36
KMI	0,39	0,27	-1,70	5,03	-0,30	1,04
Diabeto atsiradimo funkcija	0,55	0,46	-1,16	5,87	-0,42	1,20
Klasteris nr. 2						
Nėštumų skaičius	-0,33	-0,55	-1,15	2,13	-0,85	0,04
Gliukozės koncentracija	-0,38	-0,40	-3,75	1,67	-0,84	0,12
KMI	-0,30	-0,34	-2,07	2,07	-0,96	0,29
Diabeto atsiradimo funkcija	-0,42	-0,55	-1,19	1,07	-0,81	-0,10

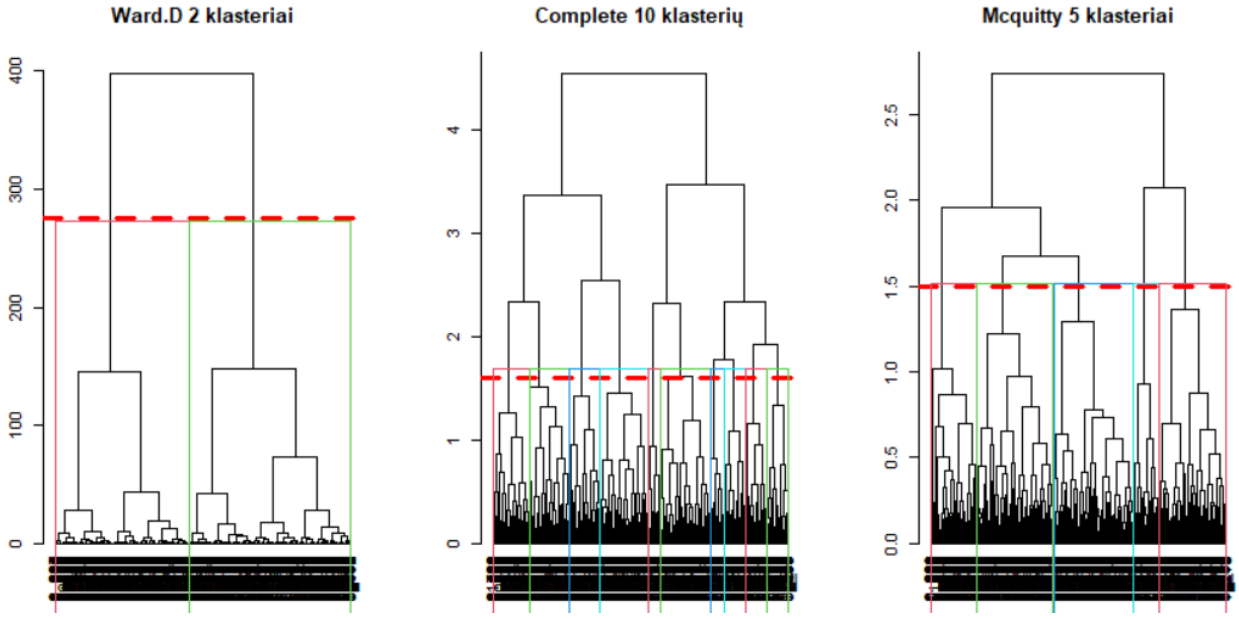
Sumažintos dimensijos

Dimensijos mažinimui buvo pritaikytas t – SNE metodas su ankstesniame darbe nustatytais geriausiais parametrais – perpleksiškumu lygiu 50 ir maksimaliu iteracijų skaičiumi lygiu 500.

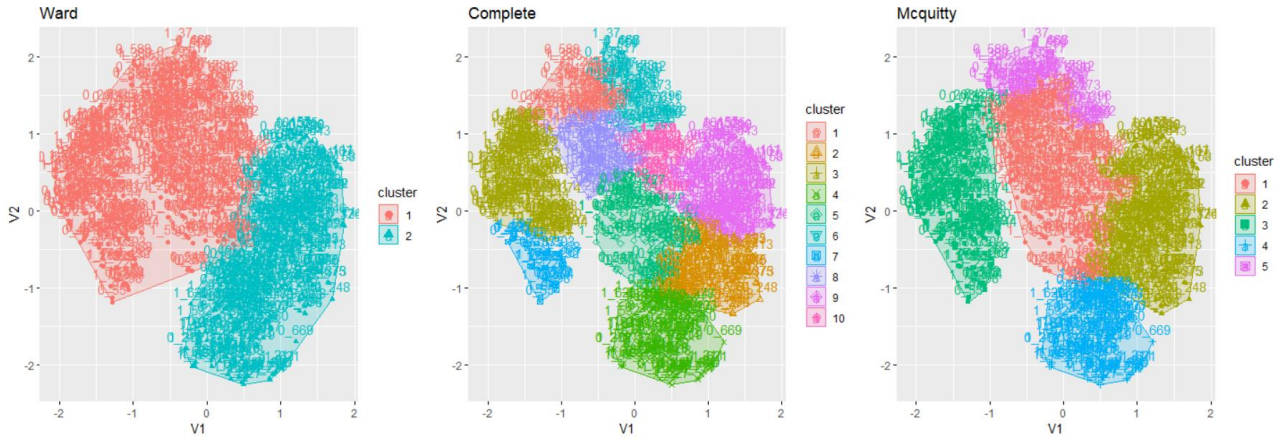


23 pav. t-SNE metodu sumažinta dimensija.

Klasterizavimas pritaikytas t-SNE metodu sumažintos dimensijos duomenų rinkiniui. Taikant Euklidinę metriką pamatėme, kad optimalus klasterių skaičius yra 2, naudojant Ward.D metodą, 10 su Complete metodu ir 5 su Mcquitty metodu.



24 pav. Dendrogramos sumažintos dimensijos duomenims su Euklidiniais atstumais.



25 pav. Klasterizavimas sumažintos dimensijos duomenims su Euklidiniais atstumais.

Galime pastebėti, jog visais atvejais klasterių persidengimas yra nedidelis.

31 lentelė. Originalus klasių pasiskirstymas.

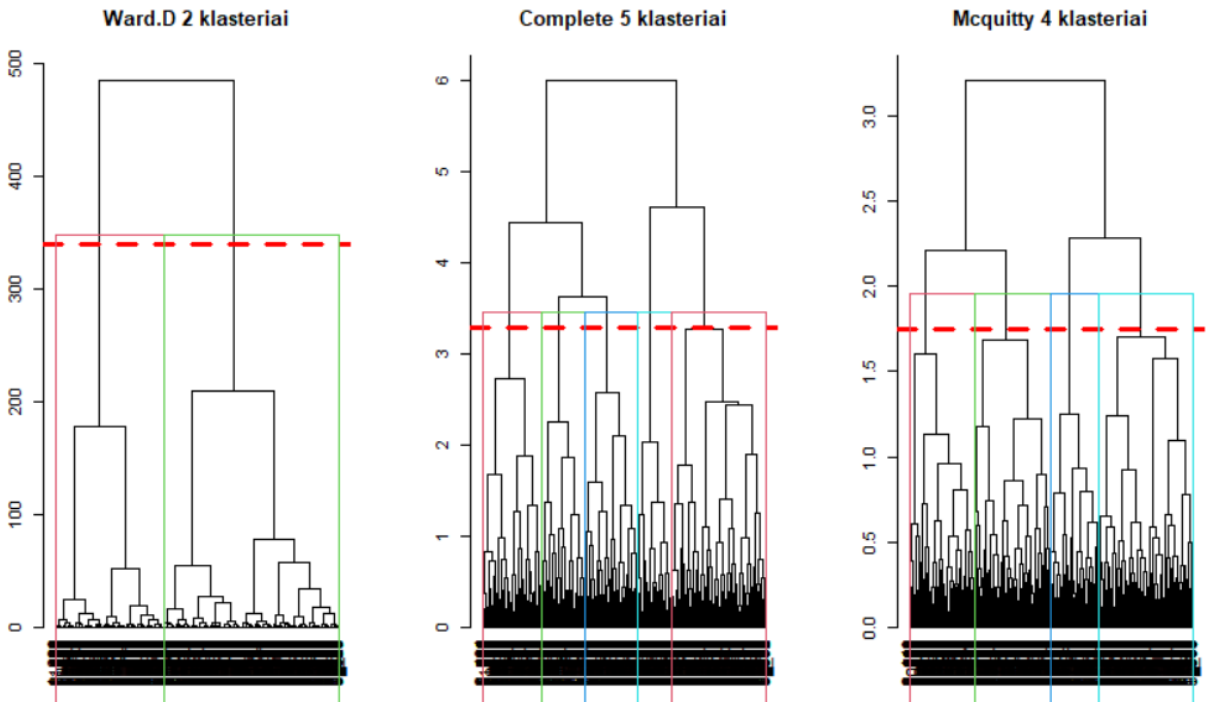
Originalus klasių pasiskirstymas	
0	1
478	251

32 lentelė. Klasterizuotų klasių pasiskirstymas su Euklidiniais atstumais naudojant Ward.D ir Mcquitty metodus

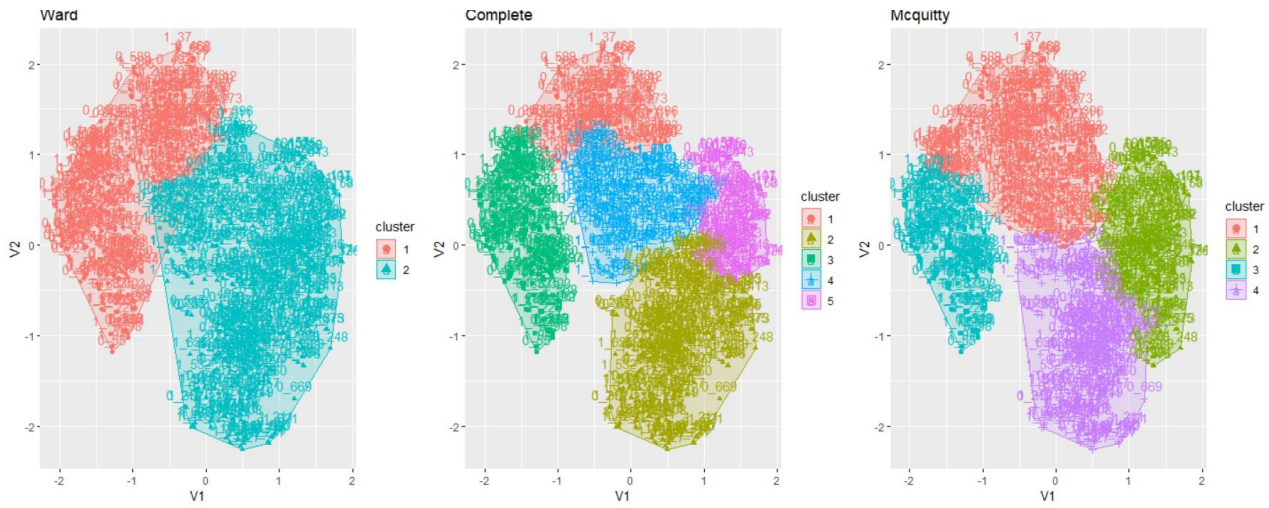
Klasterio nr.	Ward klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1
1	211	187	113	75
2	267	64	190	8
3			95	71
4			64	50
5			16	47

Galime pastebėti, jog naudojant Ward.D metodą, į 1 klasterį patenka apie 75 proc. sergančių pacientų, o į 2 apie 56 proc. visų sveikų. Mcquitty metode 2 klasteryje 96 proc. patekusių asmenų buvo sveiki.

Toliau taikome Manheteno atstumų skaičiavimo metodą. Iš dendogramų galime spręsti, jog Ward.D metodas pasiūlė 2 klasterius, Complete – 5, o Mcquitty 4 klasterius, kaip optimaliausią klasterių skaičių.



26 pav. Dendrogramos sumažintos dimensijos duomenims su Manheteno atstumais.



27 pav. Klasterizavimas sumažintos dimensijos duomenims su Manheteno atstumais.

Galime matyti, jog visuose metoduose klasteriai beveik nepersidengia.

33 lentelė. Originalus klasių pasiskirstymas.

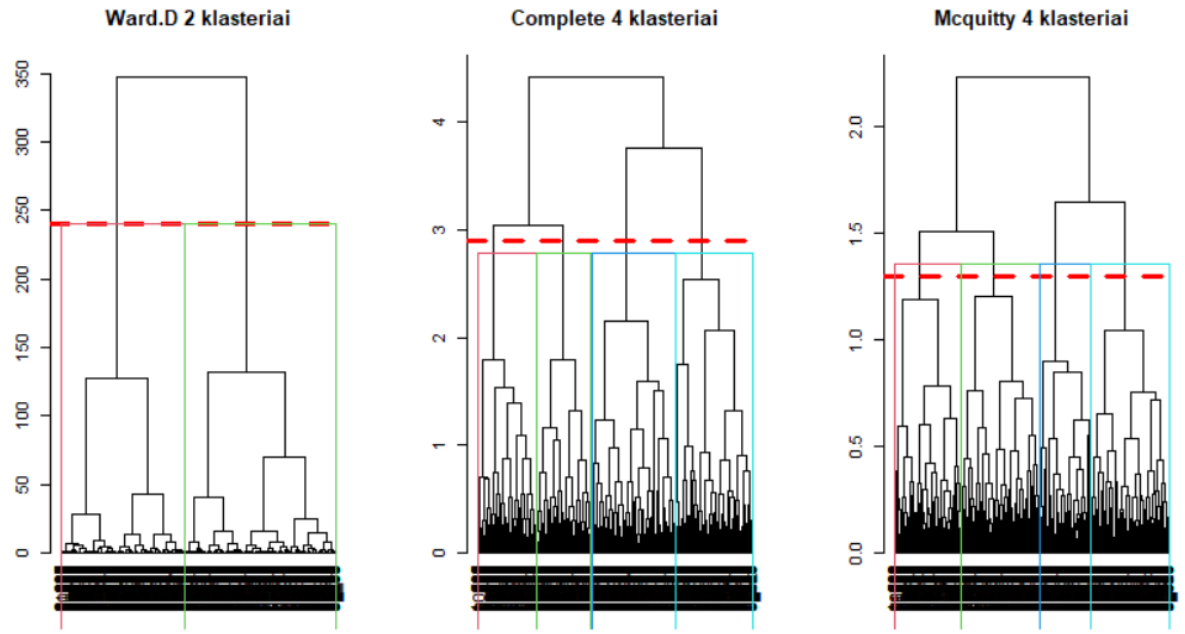
Originalus klasių pasiskirstymas	
0	1
478	251

34 lentelė. Klasterizuotų klasių pasiskirstymas su Manheteno atstumais.

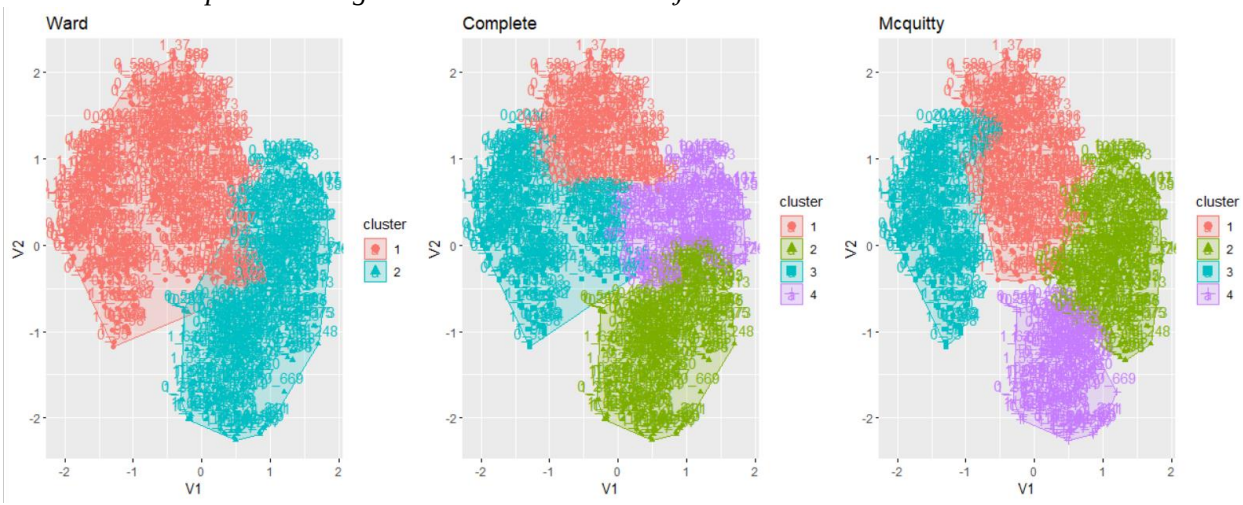
Klasterio nr.	Ward klasių klasterizavimas		Complete klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1	0	1
1	137	142	39	71	123	117
2	341	109	182	61	160	7
3			84	69	66	61
4			91	45	129	66
5			82	5		

Matome, jog taikant Ward.D metodą 70 proc. sveikų asmenų patenka į 2 klasterį. Mcquitty metodu gautą 2 klasterį sudaro 96 proc. sveikų pacientų.

Atstumų skaičiavimui dar pritaikome maksimumo metodą. Iš dendogramų galime matyti, jog Ward.D metodas siūlo 2 klasterius, Complete – 4, o Mcquitty metodas 4.



28 pav. Dendrogramos sumažintos dimensijos duomenims su maksimumo atstumais.



29 pav. Klasterizavimas sumažintos dimensijos duomenims su maksimumo atstumais.

Matome, jog klasteriai gana gerai atsiskiria, t. y. nepastebime žymaus persidengimo.

35 lentelė. Originalus klasių pasiskirstymas.

Originalus klasių pasiskirstymas	
0	1
478	251

36 lentelė. Klasterizuotų klasių pasiskirstymas su maksimumo atstumais.

Klasterio nr.	Ward klasių klasterizavimas		Complete klasių klasterizavimas		Mcquitty klasių klasterizavimas	
	0	1	0	1	0	1
1	215	187	60	95	99	109
2	263	64	164	59	200	10
3			120	86	100	76
4			134	11	79	56

Galime pastebėti, jog su Ward klasterizavimo algoritmu didžioji dalis (75 proc.) sergančių pacientų patenka į 1 klasterį.

Pagal gautus rezultatus galima matyti, kad sumažintos dimensijos duomenis vienodai gerai klasterizuoja su Euklidine ir maksimumo metrika Ward algoritmas. Maksimumo metrikos su Ward algoritmu gautos klasterių grupių aprašomosios statistikos pateiktos 27 lentelėje. Pirmos komponentės mediana žemesnė pirmame klasteryje, o antrosios komponentės – antrajame.

37 lentelė. Klasterio grupių aprašomoji statistika

Klasterizavimas reikšmingoms kovariantėms						
Klasteris nr. 1						
	Vidurkis	Mediana	Min	Max	1Q	3Q
V1	-0,69	-0,64	-2,07	0,67	-1,33	-0,12
V2	0,57	0,56	-1,18	2,16	-0,01	1,12
Klasteris nr. 2						
V1	0,85	0,87	-0,30	1,86	0,49	1,27
V2	-0,70	-0,78	-2,26	1,01	-1,32	-0,05

Apibendrinant galima teigti, jog geriausiai suklasterizavo visus požymius maksimumo metrika kartu su Ward algoritmu.

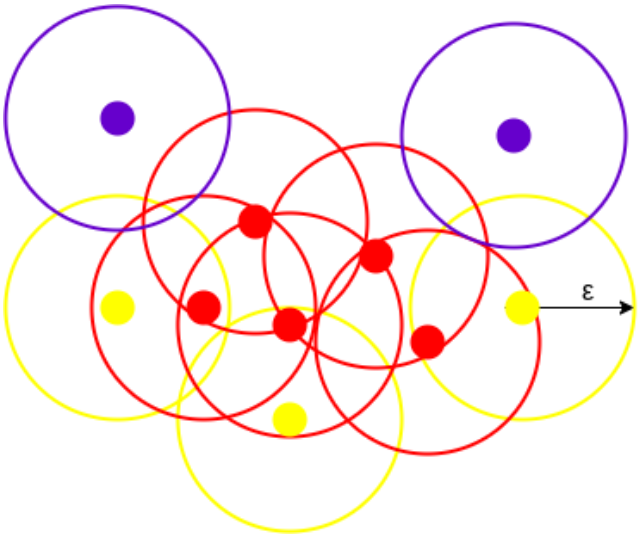
DBSCAN klasterizavimo metodas

Algoritmas taškus klasifikuoja į 3 tipus:

- pagrindiniai;
- kraštiniai;
- išskirtys.

Šis algoritmas turi du pagrindinius hiperparametrus:

- *eps* – spindulio ilgis brėžiant nuo pagrindinio.
- *minPoints* – mažiausias kaimynų skaičius apskritime.

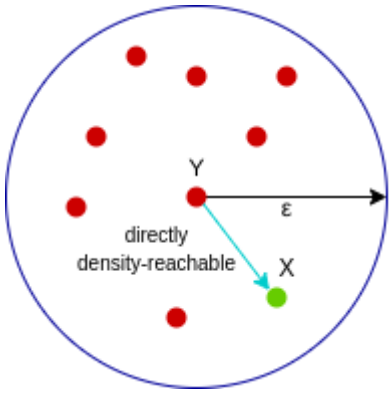


30 pav. DBSCAN algoritmo taškų skirstymas, kai kaimynų skaičius 2

(žr. 30 pav.) matome 3 skirtingomis spalvomis pažymėtus taškus: raudona, geltona ir violetinė. Šios spalvos atitinkamai žymi pagrindinius taškus, t. y. taškai, kurių apskritime yra mažiausiai 2 taškai (*minPoints*), kraštinius taškus, kurie turi taškų apskritime, bet nepakankamai, t. y. mažiau nei 2, ir taškus atsiskyrėlius, kurių apskritime nėra nei vieno taško. Apskritimų spindulys yra *eps* vertė.

Taip pat šis algoritmas taškų porą skirsto į 3 grupes:

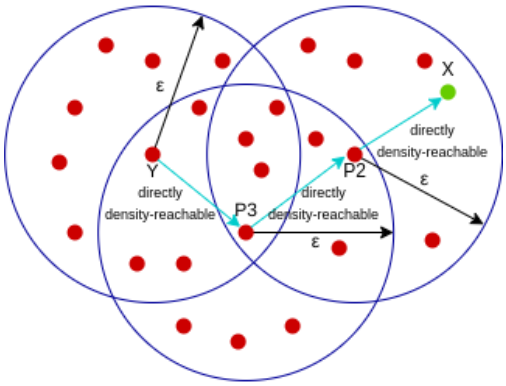
- tiesiogiai pasiekiama pagal tankumą:



31 pav. X taškas yra tiesiogiai pasiekiamas iš taško Y

(žr. 31 pav.) pateikta situacija, kai X taškas yra tiesiogiai pagal tankumą pasiekiamas iš Y taško. Būtinos sąlygos: Y turi būti pagrindinis taškas ir atstumai tarp X ir Y taškų neturi viršyti *eps* reikšmės.

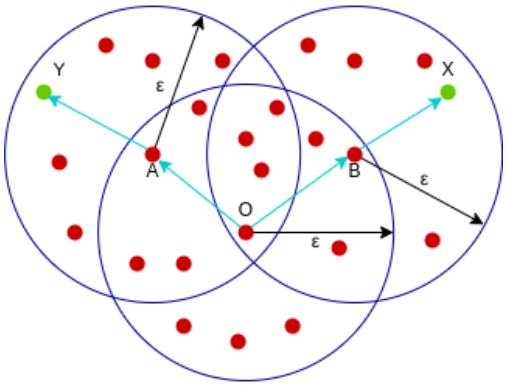
- pasiekiama pagal tankumą:



32 pav. Taškas X yra pagal tankumą pasiekiamas iš taško Y

(žr. 32 pav.) pateikta situacija, kai X taškas yra pasiekiamas pagal tankumą iš taško Y. Būtinos sąlygos: p_i taškas turi būti pasiekiamas tiesiogiai pagal tankumą iš p_{i+1} taško, čia $p_1 = X$, o $p_n = Y$.

- sujungti pagal tankumą:



33 pav. X taškas yra sujungtas pagal tankumą su Y tašku

(žr. 33 pav.) pateikta situacija, kai X taškas yra sujungtas pagal tankumą iš taško Y. Būtinos sąlygos: X ir Y taškai turi būti pasiekiami pagal tankumą iš taško O. [1]

DBSCAN algoritmo veikimas susideda iš keleto žingsnių:

1. Taškų suklasifikavimas. Algoritmas nekreipia dėmesio į triukšmą tolimesniuose etapuose.
2. Pagrindiniam taškui priskiriamas klasteris.
3. Tam pačiam klasteriui priskiriami taškai, kurie yra pagal tankumą sujungti, tačiau kraštiniai taškai nėra priskiriami. Šis etapas kartojamas, kol nebėra laisvų pagrindinių taškų.
4. Kraštiniai taškai priskiriami klasteriui, iki kurio pagrindinio taško atstumas mažiausias. [2]

Šis algoritmas turi modifikaciją, kuri atsižvelgia į galimus klasterių tankumo skirtumus, OPTICS. [3]

Silueto koeficientas bus naudojamas įvertinti klasterizavimo rezultatą:

$$s = \frac{(b-a)}{max(b-a)},$$

$$\text{čia } a$$

$$-\ atstumų\ vidurkis\ tarp\ vieno\ pasirinkto\ taško\ iki\ kitų\ to\ paties\ klasterio\ taškų,$$

$$b - atstumų\ vidurkis\ tarp\ vieno\ taško\ iki\ kitų\ kito\ arčiausio\ klasterio\ taškų.$$

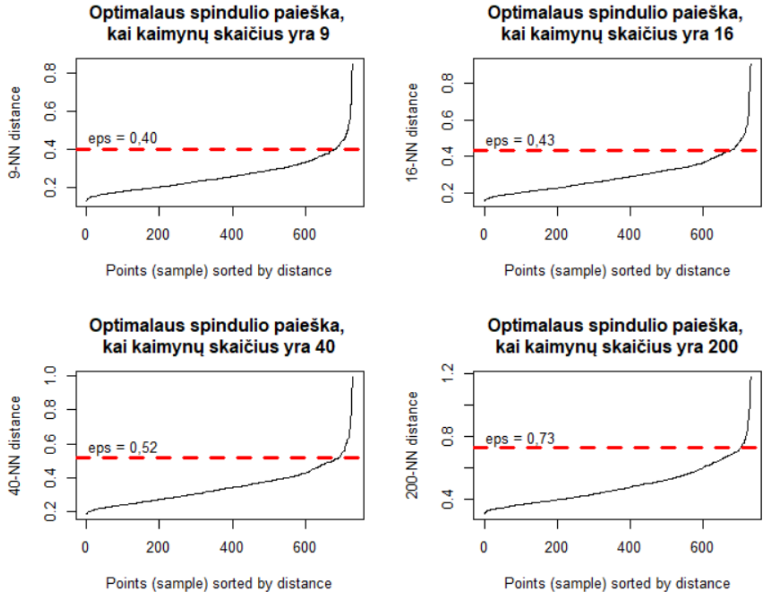
Šį koeficientą galime vertinti, kaip tankiai klasterio viduje yra išsidėstę taškai. Silueto koeficiento reikšmė kinta intervale (-1, 1), kuo reikšmė arčiau 1 tuo geriau algoritmas suklasterizavo, kuo arčiau -1, tikėtina, jog algoritmas priskyrė neteisingus taškus klasteriams, kuo arčiau 0, reiškia, kad klasteriai persidengia. [4]

Visam duomenų rinkiniui

Algoritmas buvo taikytas normuotai duomenų imčiai, taikant normavimo metodą min – max. Norint taikyti DBSCAN klasterizavimo metodą reikia surasti optimalius parametrus., t. y. spindulį ir minimalų kaimynų skaičių. Pagal rekomendacijas minimalus kaimynų skaičius turėtų būti:

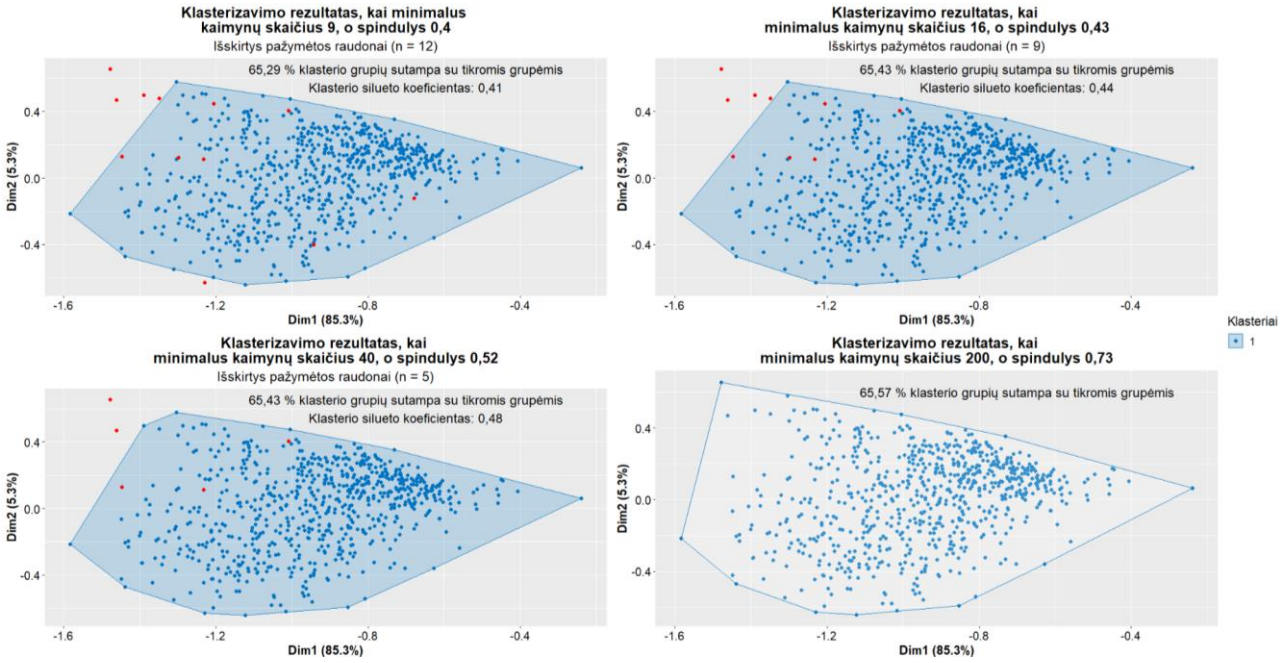
$$minPoints \geq Dim + 1, \text{čia } Dim\ duomenų\ dimensija.$$

Atlikdami tyrimą naudojome $minPoints = \{9, 16, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 250, 300, 350, 400, 450, 500\}$. Kiekvienai minimaliai kaimynų reikšmei buvo surasta atitinkama spindulio vertė – grafiko „kelio“ linkis ir bus optimali spindulio reikšmė.



34 pav. Optimalaus spindulio paieškos

Iš (žr. 34 pav.) matome, kai kaimynų skaičius yra 9, 16, 40 ir 200 optimali spindulio vertė yra atitinkamai 0,4, 0,43, 0,52, 0,73.



35 pav. Klasterizavimo rezultatai, naudojant visus požymius

Iš (35 pav.) matome, jog algoritmas išskyrė tik 1 grupė visais atvejais. Kai minimalus kaimynų skaičius buvo 200, visi taškai buvo priskirti vienam klasteriui, nebuvo rasta nei išskirčių. Matome, jog didžiausias silueto koeficientas pasiektas (0,48), kai minimalus kaimynų skaičius 40 ir spindulys 0,52. Algoritmas gerai atpažįsta išskirtis: pirmu atveju 7 pacientai, kuriuos buvome pripažinę išskirtimis, algoritmas taip pat juos pažymėjo, antru atveju tuos pačius 6 pacientus algoritmas pažymėjo atsiskyrėliais, kaip ir statistiniais metodais buvome juos pripažinę išskirtimis, trečiu atveju 4 pacientams buvo priskirta kategorija išskirtis, kaip ir mes buvome anksčiau tiems patiems pacientams nustatę.

Su likusiomis parametrų reikšmėmis nebuvo gauta geresnių rezultatų.

Toliau pateiksime klasterio išskirčių aprašomąją statistiką, kai minimalus kaimynų skaičius 40 ir spindulys 0,52, palyginimui su mūsų atrastų išskirčių aprašomąja statistika.

38 lentelė. Išskirčių aprašomosios statistikos

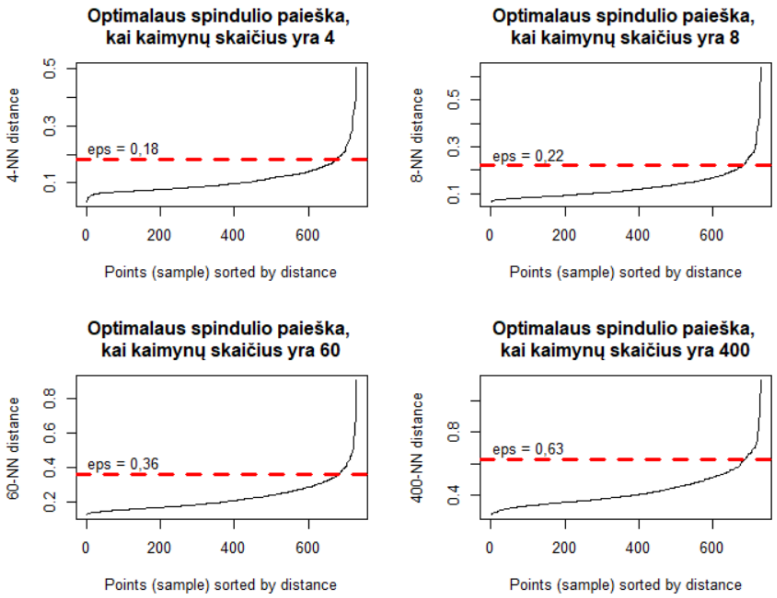
Klasterio rastos išskirtys (n = 5)				
	Vidurkis	Mediana	Min	Max
Nėštumų skaičius	1,40	1,00	0,0	4,00
Gliukozės koncentracija kraujyje	180,00	189,00	137,00	197,00
Kraujo spaudimas	63,60	70,00	40,00	78,00
Odos storis	51,80	39,00	23,00	99,00
Insulino koncentracija	354,40	168,00	0,000	846,00
KMI	40,8	36,70	30,1	59,40
Diabeto atsiradimo funkcija	1,60	2,29	0,400	2,42
Amžius	42,00	33,00	25,00	62,00
Tikros išskirtys (n = 16)				
Nėštumų skaičius	2,00	1,50	0,00	9,00
Gliukozės koncentracija kraujyje	147,20	149,50	82,00	197,00
Kraujo spaudimas	73,38	75,00	40,00	110,00
Odos storis	34,25	35,00	0,00	63,00
Insulino koncentracija	340,40	357,5	0,00	846,00
KMI	39,58	39,05	24,20	67,10
Diabeto atsiradimo funkcija	1,22	1,14	0,13	2,42
Nėštumų skaičius	33,00	27,00	21,00	81,00

Iš (žr. 38 lentelė) matome, jog klasteris glukozės, odos storio, diabeto atsiradimo funkcijos ir amžiaus kovariantėms naudojo didesnės vertės slenksčius išskirtims aptikti nei mes su statistiniais metodais, nes šių kovariančių medianos reikšmė didesnė nei tikrųjų išskirčių, o likusioms kovariantėms naudojo priešingai – mažesnės vertės slenksčius.

Grupių aprašomosios statistikos nelyginsime, nes algoritmas išskyrė tik vieną grupę.

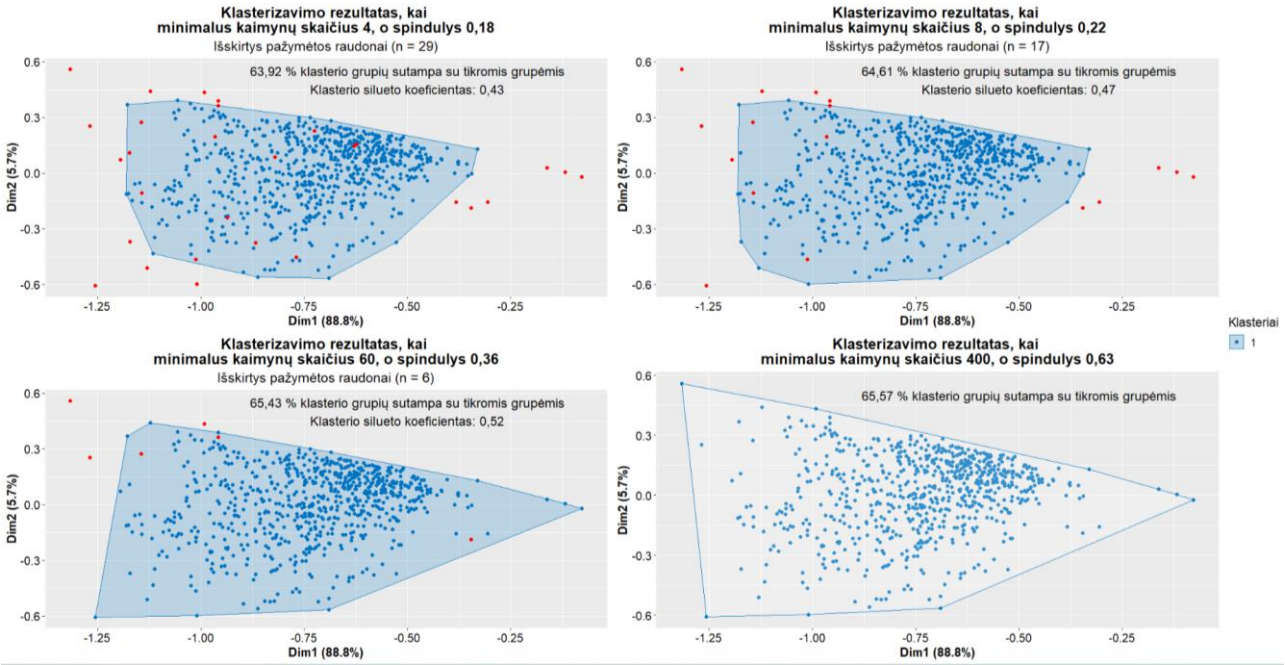
Reikšmingoms kovariantėms

Klasterizavimą vykdėme su pasirinktomis reikšmingomis kovariantėmis iš logistinio modelio, kurios yra: nėštumų skaičius, gliukozės koncentracija kraujyje, KMI indeksas ir diabeto atsiradimo funkcija. Algoritmas buvo taikytas normuotai duomenų imčiai, taikant normavimo metodą min – max. DBSCAN taikymą pradėjome nuo optimalių parametrų paieškos. Atlikdami tyrimą naudojome $minPoints = \{4, 5, 6, 7, 8, 16, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 250, 300, 350, 400, 450, 500\}$. Kiekvienai minimaliai kaimynų reikšmei buvo surasta atitinkama spindulio vertė – grafiko „kelio“ linkis ir buvo optimali spindulio reikšmė.



36 pav. Optimalaus spindulio paieškos

Iš (žr. 36 pav.) matome, kad esant kaimynų skaičiui 4, 8, 60, 400 optimalus spindulys atitinkamai yra 0,18, 0,22, 0,36 ir 0,63. Šiems parametrų rinkiniams pateiksime klasterizavimo rezultatus.



37 pav. Klasterizavimo rezultatai, kai naudojami 4 požymiai

Iš (žr. 37 pav.) matome, jog nė viena parametrų kombinacija neišskyrė dviejų ar daugiau klasterių. Kai minimalus kaimynų skaičius buvo 400 ir spindulio vertė 0,63, visi taškai buvo priskirti vienam klasteriui, algoritmas neberado nei išskirčių. Didžiausia silueto koeficiento reikšmė (0,52) buvo pasiekta, kai minimalus kaimynų skaičius buvo 60, o spindulys 0,36. Algoritmas neblogai pasirodė atpažįstant išskirtis: pirmajame grafike iš 29 algoritmo pripažintų išskirčių 8 tos pačios išskirtis buvo pripažintos naudojantis įprastais statistiniais metodais,

antrajame grafike 6 tas pačias išskirtis rado algoritmas kaip ir su įprastais statistiniais metodais, o trečiajame grafike 5 iš 6 algoritmo rastų išskirčių buvo tos pačios, kaip ir mūsų anksčiau rastos.

Su likusiomis parametrų reikšmėmis nebuvo gauta geresnių rezultatų.

Toliau pateiksime klasterio išskirčių aprašomąją statistiką, kai minimalus kaimynų skaičius 60 ir spindulys 0,36, palyginimui su mūsų atrastų išskirčių aprašomąja statistika nenormuotiems duomenims.

39 lentelė. Išskirčių aprašomosios statistikos palyginimas

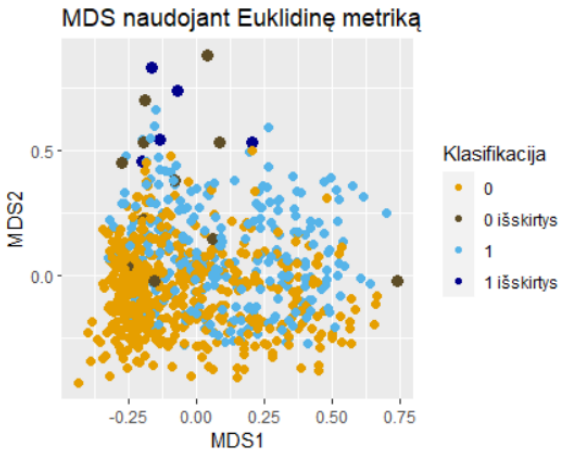
Klasterio rastos išskirtys (n = 6)				
	Vidurkis	Mediana	Min	Max
Nėštumų skaičius	2,17	1,50	0,00	6,00
Gliukozės koncentracija kraujyje	136,00	155,00	0,00	197,00
KMI	47,28	41,50	36,70	67,10
Diabeto atsiradimo funkcija	1,70	2,21	0,32	2,42
Tikros išskirtys (n = 16)				
Nėštumų skaičius	2,00	1,50	0,00	9,00
Gliukozės koncentracija kraujyje	147,20	149,50	82,00	197,00
KMI	39,58	39,05	24,20	67,10
Diabeto atsiradimo funkcija	1,22	1,14	0,13	2,42

Iš (žr. 39 lentelė) matome, jog nėštumų skaičiaus mediana sutampa abiejuose išskirčių rinkiniuose. Gliukozės koncentracijai, KMI indeksui, diabeto atsiradimo funkcijai medianos reikšmė didesnė algoritmo rastoms išskirtims, atitinkamai 155,00 ir 149,50, 41,50 ir 39,05, 2,21 ir 1,14. Tai reiškia, jog algoritmas šioms kovariantėms naudojo didesnes slenksčių reikšmes palyginus su mūsų statistiniais metodais.

Grupių aprašomosios statistikos nelyginsime, nes algoritmas išskyrė tik vieną grupę.

Sumažintos dimensijos

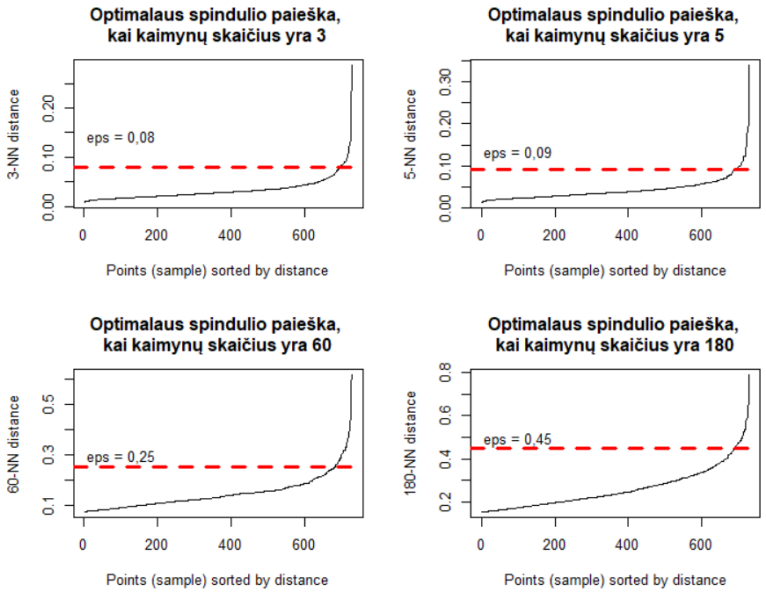
Dimensijos mažinimui pasirinkome MDS metodą su Euklidine metrika.



38 pav. Dimensijos mažinimo rezultatas, naudojant MDS metodą su Euklidine metrika

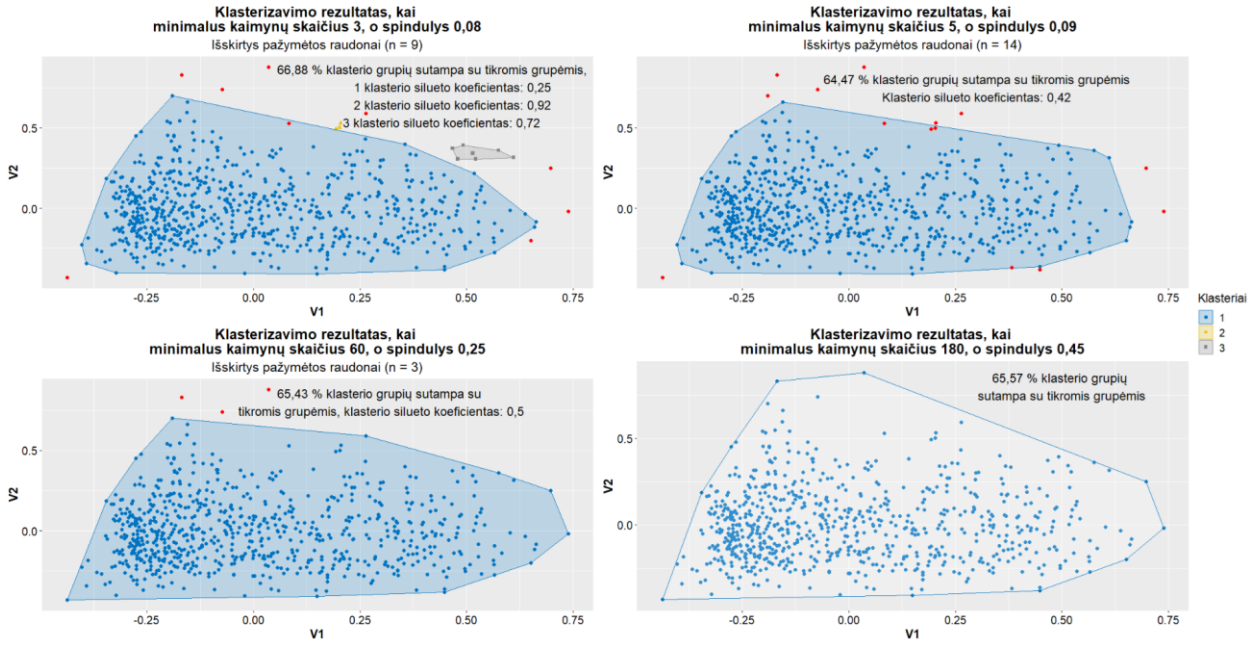
Duomenų vaizdas po dimensijos mažinimo yra pateikiamas (žr. 38 pav.).

Algoritmo taikymą pradėjome nuo optimalių parametrų paieškos. Atlikdami tyrimą naudojome $minPoints = \{3, 5, 7, 9, 11, 13, 15, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 300, 400, 500\}$. Kiekvienai minimaliai kaimynų reikšmei buvo surasta atitinkama spindulio vertė – grafiko „kelio“ linkis ir buvo optimali spindulio reikšmė.



39 pav. Optimalių spindulių paieškos, kai naudojama sumažintos dimensijos duomenys

Iš (žr. 39 pav.) matome, jog esant kaimynų skaičiui 3, 5, 60, 180 atitinkamai optimaliausia spindulio vertė yra 0,08, 0,09, 0,25 ir 0,45. Šiems parametrų rinkiniams pateiksime klasterizavimo rezultatus.



40 pav. Klasterizavimo rezultatai, kai naudojami duomenys po dimensijos mažinimo

Iš (žr. 40 pav.) matome, jog esant minimaliam kaimynų skaičiui 3 ir spindulio reikšmei 0,08 yra išskiriami 3 klasteriai - nors ir dominuoja mėlynasis klasteris, tačiau algoritmas sugebėjo išskirti dar 2 grupes. Taip pat šiam parametrui rinkiniui pasiektas geriausias klasių sutapimo procentas 66,88, t. y. kiek klasterių priskirtų etikečių sutapo su turimomis. Išskirtas geltonam klasteriui buvo priskirti 3 pacientai, o pilkam 7. Kitiems parametrų rinkiniams buvo rastas tik 1 klasteris. Esant minimaliam kaimynų skaičiui 180 ir spinduliui 0,45 visi taškai buvo priskirti klasteriui, t. y. nebuvo rasta nei išskirčių. Algoritmas pasižymėjo geru išskirčių atpažinimu: pirmuoju atveju 5 iš 9 pripažintos išskirtys buvo identifikuotos su statistiniais metodais, antruoju atveju 7 iš 14 algoritmo rastų išskirčių taip pat buvo rastos su statistiniais metodais, o trečiuoju – 100 proc. priskirtos išskirtys buvo rastos su statistiniais metodais.

Toliau pateiksime klasterio išskirčių aprašomąją statistiką, kai minimalus kaimynų skaičius 3 ir spindulys 0,08, palyginimui su mūsų atrastų išskirčių aprašomąja statistika. Kadangi MDS algoritmas neįvardina, kokias kovariantes palieka, aprašomąją statistiką pateiksime visoms kovariantėms.

40 lentelė. Išskirčių palyginimas, naudojant sumažintos dimensijos duomenis

Klasterio rastos išskirtys (n = 9)				
Nėštumų skaičius	Vidurkis	Mediana	Min	Max
Gliukozės koncentracija kraujyje	5,89	4,00	0,00	17,00
Kraujo spaudimas	146,67	163,00	0,00	197,00
Odos storis	72,89	74,00	48,00	84,00
Insulino koncentracija	44,89	41,00	17,00	99,00
KMI	215,78	60,00	0,00	744,00
Diabeto atsiradimo funkcija	36,21	36,70	24,70	59,40
Amžius	1,08	0,62	0,14	2,42
Nėštumų skaičius	43,00	34,00	22,00	81,00
Tikros išskirtys (n = 16)				
Nėštumų skaičius	2,00	1,50	0,00	9,00
Gliukozės koncentracija kraujyje	147,20	149,50	82,00	197,00
Kraujo spaudimas	73,38	75,00	40,00	110,00
Odos storis	34,25	35,00	0,00	63,00
Insulino koncentracija	340,40	357,50	0,00	846,00
KMI	39,58	39,05	24,20	67,10
Diabeto atsiradimo funkcija	1,22	1,14	0,13	2,42
Amžius	33,00	27,00	21,00	81,00

Iš (žr. 40 lentelė) matome, jog nėštumų skaičiui, gliukozės koncentracijai kraujyje, odos storiui ir amžiui medianos reikšmė didesnė DBSCAN algoritmo rastoms išskirtims, o likusioms kovariantėms priešingai mažesnė.

Taip pat pateiksime klasterio gautų grupių aprašomąją statistiką.

41 lentelė. Klasterio grupių aprašomoji statistika

1 grupė (n = 710)						
Nėštumų skaičius	Vidurkis	Mediana	Min	Max	Standartinis nuokrypis	Plotis
Gliukozės koncentracija kraujyje	3,79	3,00	0,00	15,00	3,28	15,00
Kraujo spaudimas	120,00	115,50	0,00	199,00	31,22	199,00
Odos storis	72,25	72,00	24,00	122,00	12,35	98,00
Insulino koncentracija	21,01	23,00	0,00	60,00	15,36	60,00
KMI	78,29	44,00	0,00	680,00	105,56	680,00
Diabeto atsiradimo funkcija	32,36	32,15	18,20	67,10	6,81	48,90
Amžius	0,46	0,37	0,08	2,29	0,31	2,21
Nėštumų skaičius	32,89	29,00	21,00	70,00	11,34	49,00
2 grupė (n = 3)						
Nėštumų skaičius	1,00	1,00	0,00	2,00	1,00	2,00
Gliukozės koncentracija kraujyje	186,33	189,00	173,00	197,00	12,22	24,00
Kraujo spaudimas	69,33	70,00	60,00	78,00	9,02	18,00
Odos storis	33,33	32,00	23,00	45,00	11,06	22,00
Insulino koncentracija	551,33	543,00	265,00	846,00	290,59	581,00
KMI	35,70	30,50	30,10	46,50	9,36	16,40
Diabeto atsiradimo funkcija	0,57	0,40	0,16	1,16	0,52	1,00
Amžius	56,67	58,00	53,00	59,00	3,21	6,00
3 grupė (n = 7)						
Nėštumų skaičius	9,71	9,00	8,00	13,00	2,06	5,00
Gliukozės koncentracija kraujyje	166,43	171,00	140,00	196,00	20,07	56,00
Kraujo spaudimas	85,43	84,00	68,00	110,00	13,15	42,00
Odos storis	35,86	36,00	24,00	48,00	8,07	24,00
Insulino koncentracija	288,14	280,00	140,00	495,00	108,96	355,00
KMI	37,73	37,60	30,10	45,40	4,89	15,30

Diabeto atsiradimo funkcija	0,73	0,62	0,47	1,17	0,26	0,71
Amžius	53,86	57,00	39,00	60,00	7,20	21,00

Iš (žr. 41 lentelė) matome, jog 1 – jame klasteryje dažniausiai medianos reikšmė mažiausia iš visų klasterių, t. y. gliukozės koncentracijos kraujyje, odos storio, insulino kiekio, diabeto atsiradimo funkcijos ir amžiaus kovariantėms mažiausia medianos reikšmė. Galėtumėme įtarti, jog vyrauja sveiki pacientai. 3 – jame klasteryje vyrauja stebiniai, kurių medianos reikšmė didžiausia iš visų kitų grupių nėštumo skaičiaus, kraujo spaudimo, odos storio, KMI ir diabeto atsiradimo funkcijos kovariantėms. Galėtumėme įtarti, jog čia vyrauja sergantys pacientai, o 2 klasterį traktuoti kaip pereinamąjį, t. y. įtartini pacientai. Tačiau šiuo klasterizavimu negalėtumėme remtis tolimesniuose tyrimuose, nes vyrauja viena grupė sudaranti 99 proc. stebėjimų.

Galime daryti išvadą, jog DBSCAN klasterizavimo metodas neparodė gerų rezultatų nagrinėjamiems duomenims, tačiau neblogai atpažino išskirtis.

IŠVADOS

Klasterizavimo metodai buvo pritaikyti visiems požymiams, reikšmingoms kovariantėms (nėštumų skaičiui, gliukozės koncentracijai kraujyje, KMI indeksui ir diabeto atsiradimo funkcijai) bei sumažinus dimensiją iki 2.

Naudojant *k – vidurkių* metodą, gavome, jog optimalus klasterių skaičius tiek visų kovariančių, tiek 4 kovariančių, tiek naudojant MDS dimensijų mažinimo algoritmą su elbow metodu yra 3, vidutinio silueto – 2. Naudojant empirinio vidurkio metodą - geriausiai 2 klasteriai buvo atskirti su 4 kovariantėmis, jų R^2 buvo 61,1%. Geriausiai klasterizavimas pasirodė naudojant reikšmingas kovariantes.

Klasterizavimui pritaikius hierarchinį algoritmą dendogramos neparodė vieningos reikšmės klasterių skaičiui nusakyti. Skirtingiems duomenų rinkiniams buvo pritaikyti 3 atstumų skaičiavimo metodai: Euklidinis, Manheteno, maksimumo ir 3 klasterizavimo metodai: Ward.D, Complete ir Mcquitty. Geriausias rezultatas buvo pasiektas naudojant visus požymius su maksimumo metrika kartu su Ward.D algoritmu - apie 98 proc. sergančių pacientų pateko į 1 klasterį, o apie 70 proc. sveikų į 2 klasterį.

DBSCAN algoritmas visiems požymiams ir reikšmingiems išskyrė tik vieną klasterį visiems tirtiems parametrų rinkiniams bei atpažino išskirtis. Nagrinėjant MDS algoritmu sumažintos dimensijos duomenis algoritmas su minimaliu kaimynų skaičiumi 3 ir spinduliu 0,08, išskyrė 3 klasterius, tačiau 2 iš jų sudarė tik apie 2 proc. stebėjimų. Šis algoritmas neparodė gerų klasterizavimo rezultatų nagrinėjamai duomenų imčiai, tačiau neblogai identifikavo išskirtis.

Geriausias rezultatas lyginant visus klasterizavimo metodus buvo pasiektas naudojant visus požymius su maksimumo metrika kartu su Ward.D algoritmu - apie 98 proc. sergančių pacientų pateko į 1 klasterį, o apie 70 proc. sveikų į 2 klasterį. Antras geriausias rezultatas pasiektas, naudojant reikšmingas kovariantes ir taikant k – means algoritmą su klasterių skaičiumi 2: 84 proc. nesergančių žmonių pateko į 1 klasterį, tačiau šiame klasteryje taip pat daug (67 proc.) sergančių pacientų.

LITERATŪRA

[1] Sharma, A., „How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning“, 2020: https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/#:~:text=DBSCAN%20is%20a%20density%2Dbased,points%20into%20a%20single%20cluster

[2] Thailappan, D., „Understand The DBSCAN Clustering Algorithm!“, 2022: https://www.analyticsvidhya.com/blog/2021/06/understand-the-dbscan-clustering-algorithm/

[3] Yufeng, „Understanding OPTICS and Implementation with Python“, 2022: https://towardsdatascience.com/understanding-optics-and-implementation-with-python-143572abdfb6#:~:text=The%20OPTICS%20Algorithm&text=In%20OPTICS%20each%20point%20is,of%20p%2C%20whichever%20is%20bigger

[4] Mulin, T., „DBSCAN — Overview, Example, & Evaluation“, 2020: https://medium.com/@tarammullin/dbscan-2788cfce9389#:~:text=DBSCAN%20has%20two%20parameters.,to%20form%20a%20distinct%20cluster

[5] https://uc-r.github.io/hc_clustering

[6] https://uc-r.github.io/kmeans_clustering#gap