
VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFOMATIKOS FAKULTETAS

DUOMENŲ MOKSLO BAKALAURAS

# DIMENSIJOS MAŽINIMAS

Ataskaita

Atliko: Simona Gelžinytė,

Ugnė Kniukštaitė, Rugilė Bagdonaitė

duomenų mokslas 3 k.

Vilnius, 2023

# Turinys

# ĮVADAS

## Tikslas

Panaudoti dimensijos mažinimo metodus daugiamačių duomenų vizualizavimui, ištirti metodų galimybes.

## Uždaviniai

1. Pateikti duomenų aibės aprašomąją statistiką, aprašyti duomenų aibės specifiką.
2. Ištirti išskirtis bei apžvelgti požymių koreliacijas.
3. Sunormuoti duomenų aibę pagal vidurkį ir dispersiją.
4. Sumažinti duomenų aibės dimensiją iki dim = 2, naudojant tris pasirinktus dimensijos mažinimo metodus. Tai padaryti su normuota ir nenormuota duomenų aibe.
5. Ištirti, kaip keičiasi vizualizavimo rezultatai, keičiant algoritmų parametrus.
6. Vizualizuoti rezultatus naudojant taškinius grafikus, klases arba klasterius atskiriant skirtingomis spalvomis.
7. Įvertinti gautus rezultatus, padaryti išvadas, kuris metodas geriau atvaizduoja rezultatą.

## Duomenys

Pasirinktas duomenų rinkinys apie diabetą[1]. Duomenų rinkinyje pateikti medicininiai požymiai, padedantys nustatyti ar pacientas serga diabetu. Buvo tirtos tik moterys, ne jaunesnės nei 21 metų. Priklausomas kintamasis – rezultatas, rodantis ar pacientas serga diabetu (1 – serga, 0 – neserga) ir 8 kovariantės:

- Nėštumas – nėštumų skaičius;
- Gliukozė – gliukozės koncentracija plazmoje;
- Kraujo spaudimas – diastolinis kraujo spaudimas (mm Hg);
- Odos storis – tricepso odos raukšlės storis (mm);
- Insulinas – 2 valandų serumo insulinas (mU / ml);
- KMI – kūno masės indeksas;
- Diabeto kilmės funkcija;
- Amžius.

Duomenys paimti iš [1] šaltinio.

Iš viso yra 768 stebėjimai. Duomenų aibėje nėra praleistų reikšmių, tačiau pastebėjome, kad kai kurie įrašai neatitinka logiškos kintamųjų skalės (pvz. kraujo spaudimas ar KMI lygūs 0), todėl darome 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 padaryti prielaida, jog yra teisingi.

*1 lentelė. Duomenų tipai ir skalės*

| | Pregnancies | Glucose | Blood Pressure | Skin Thickness | |
|---|---|---|---|---|---|
| Duomenų tipas | Diskretieji | Diskretieji | Diskretieji | Diskretieji | |
| Skalė | Intervalų | Intervalų | Intervalų | Intervalų | |
| | Insulin | BMI | Diabetes pedigree Function | Age | Outcome |
| Duomenų tipas | Diskretieji | Tolydieji | Tolydieji | Diskretieji | Dvinariai |
| Skalė | Intervalų | Santykių | Santykių | Intervalų | Nominalinė |

(1 lentelė) yra pateikta požymių tipas ir skalė.

# PIRMINIS DUOMENŲ APDOROJIMAS

## Pradinė aprašomoji statistika

Patikriname, ar kuri nors iš ligos statusą nurodančių kategorijų įrašai nesudaro daugiau nei 80 % duomenų rinkinio.

*2 lentelė. Procentinė dalis*

| Neserga | Serga |
|---|---|
| 0.6556927 | 0.3443073 |

Iš rezultatų galima matyti, kad nesergantys asmenys sudaro 66 % rinkinio ir sergantys – 34 %, todėl galime teigti, jog duomenys yra tinkami.

Skaitiniams rodikliams apskaičiuotos pagrindinės aprašomosios statistikos charakteristikos (standartinis nuokrypis, vidurkis, mediana, mažiausia reikšmė (min), didžiausia reikšmė (max), 1 ir 3 kvartilis).

*3 lentelė. Pradinė duomenų aprašomoji statistika*

| | stand. nuokr. | vidurkis | mediana | min | max | $Q_1$ | $Q_3$ |
|---|---|---|---|---|---|---|---|
| Pregnancies | 3,357 | 4 | 3 | 0 | 17 | 1 | 6 |
| Glucose | 32,255 | 121 | 117 | 0 | 199 | 99 | 141 |
| Blood Pressure | 12,376 | 72 | 72 | 24 | 122 | 64 | 80 |
| Skin Thickness | 15,708 | 22 | 24 | 0 | 99 | 0 | 33 |
| Insulin | 116,803 | 84 | 46 | 0 | 846 | 0 | 130 |
| BMI | 6,885 | 32,47 | 32,4 | 18,2 | 67,1 | 27,5 | 36,6 |
| Diabetes Pedigree Function | 0,332 | 0,4741 | 0,378 | 0,078 | 2,42 | 0,245 | 0,627 |
| Age | 11,753 | 33 | 29 | 21 | 81 | 24 | 41 |

Tos pačios charakteristikos apskaičiuotos ir skirtingoms pacientų grupėms, t.y. kai diagnozuotas diabetas ir, kai ne (4 lentelė - 5 lentelė). Pasirinkus lyginimo charakteristiką – medianą, matome, jog pacientės, kurioms diagnozuotas diabetas, turėjusios daugiau nėštumų (medianinė reikšmė = 5), jų gliukozės koncentracija plazmoje didesnė (medianinė reikšmė = 140), aukštesnis kraujo spaudimas (medianinė reikšmė = 74), storesnė tricepso odos raukšlė (medianinė

reikšmė = 28), didesnis insulino kiekis (medianinė reikšmė = 58), didesnis KMI (medianinė reikšmė = 34,3) bei yra vyresnės (medianinė reikšmė = 36).

Group 1

*4 lentelė. Group 1 duomenų aprašomoji statistika*

| | stand. nuokr. | vidurkis | mediana | min | max |
|---|---|---|---|---|---|
| Pregnancies | 3,69 | 5 | 5 | 0 | 17 |
| Glucose | 32,52 | 141 | 140 | 0 | 199 |
| Blood Pressure | 12,25 | 75 | 74 | 30 | 114 |
| Skin Thickness | 17,33 | 24 | 28 | 0 | 60 |
| Insulin | 140,75 | 107 | 58 | 0 | 846 |
| BMI | 6,58 | 35,35 | 34,3 | 22,9 | 67,1 |
| Diabetes Pedigree Function | 0,38 | 0,56 | 0,45 | 0,09 | 2,42 |
| Age | 11,08 | 37 | 36 | 21 | 70 |

*Group 0*

*5 lentelė. Group 0 duomenų aprašomoji statistika*

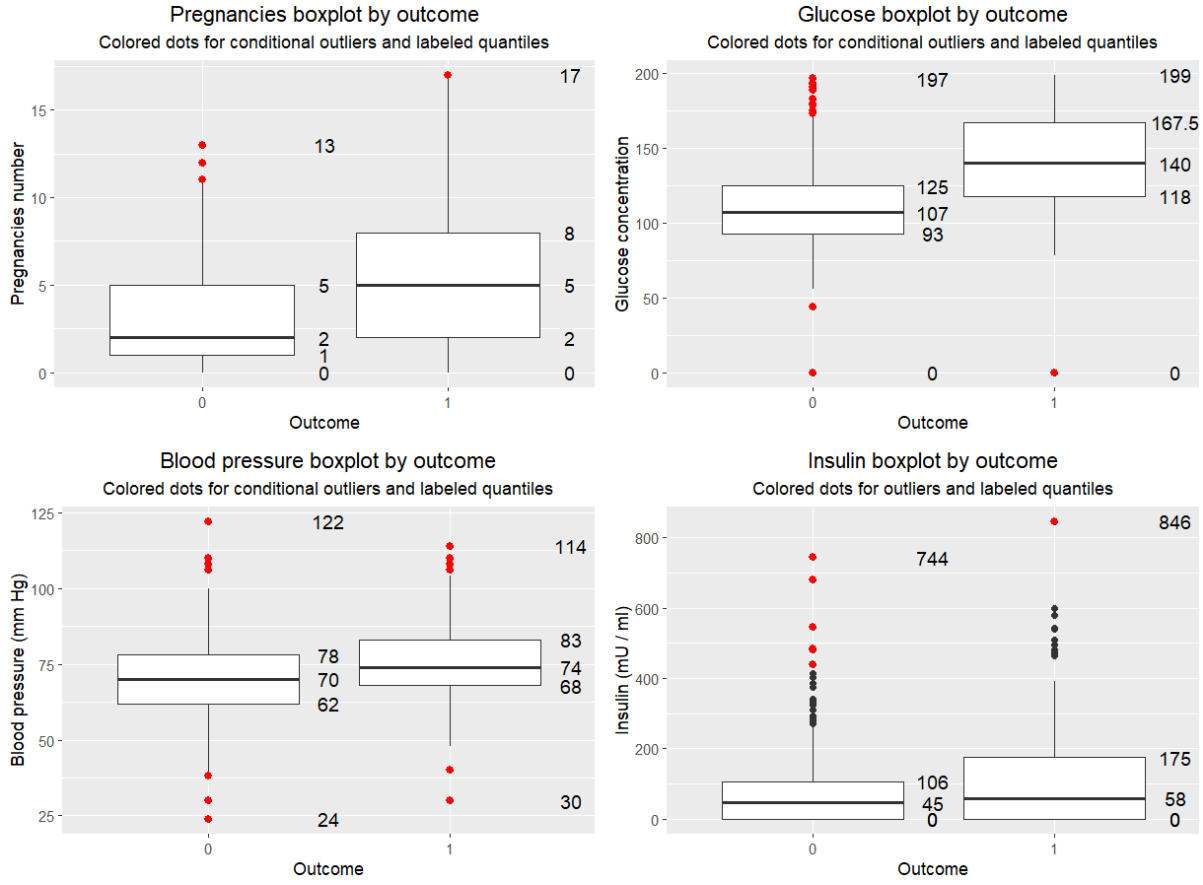
| | stand. nuokr. | vidurkis | mediana | min | max |
|---|---|---|---|---|---|
| Pregnancies | 3,03 | 3 | 2 | 0 | 13 |
| Glucose | 26,43 | 110 | 107 | 0 | 197 |
| Blood Pressure | 12,18 | 71 | 70 | 24 | 122 |
| Skin Thickness | 14,69 | 20 | 22 | 0 | 60 |
| Insulin | 100,04 | 72 | 45 | 0 | 744 |
| BMI | 6,56 | 30,96 | 30,4 | 18,2 | 57,3 |
| Diabetes Pedigree Function | 0,3 | 0,43 | 0,34 | 0,08 | 2,33 |
| Age | 11,55 | 31 | 27 | 21 | 81 |

## Išskirčių analizė

Norint atpažinti išskirtis duomenų aibėje, pasinaudojome statistiniais išskirčių apibrėžimais:

- Sąlyginė išskirtis identifikuojama, kai stebėjimo reikšmė yra tarp vidinio ir išorinio barjero, t. y. kai stebėjimas pakliūva į intervalą $(Q_1 - 3 \times (Q_3 - Q_1); Q_1 - 1{,}5 \times (Q_3 - Q_1)]$ arba $[Q_1 + 1{,}5 \times (Q_3 - Q_1); Q_1 + 3 \times (Q_3 - Q_1))$.
- Išskirtis nustatoma, kai stebėjimas yra už išorinio barjero ribos, t. y. reikšmė $< Q_1 - 3 \times (Q_3 - Q_1)$ arba $> Q_1 + 3 \times (Q_3 - Q_1)$.

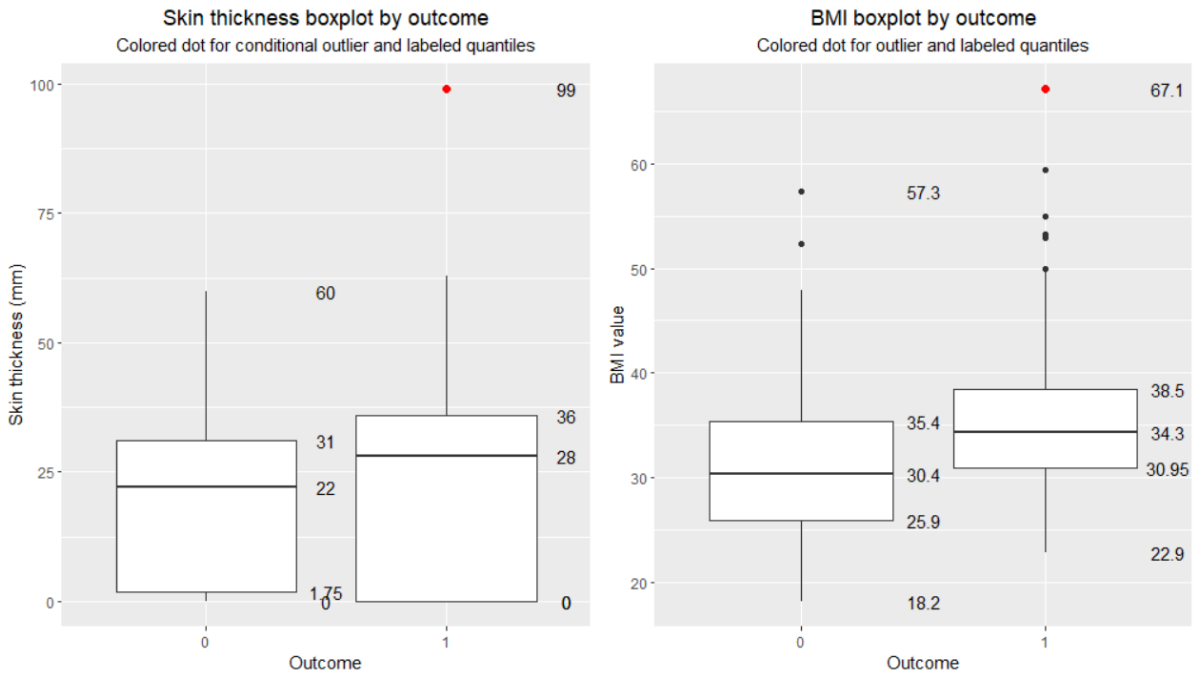
Išskirčių ieškojome visiems požymiams. Iš viso buvo rasta 16 išskirčių ir 111 sąlyginių išskirčių iš visų požymių. Buvo nuspręsta išskirčių nešalinti ir nemodifikuoti, bet jas kaupti naujoje lentelėje, kad būtų galima pažiūrėti, ar dimensijos mažinimo algoritmai reaguos ir kažkaip atskirs stebėjimus atsiskyrėlius.




*1 pav. Stačiakampių diagramų panelė, tirianti gliukozę, kaujo spaudimą, insuliną ir nėštumų skaičių*

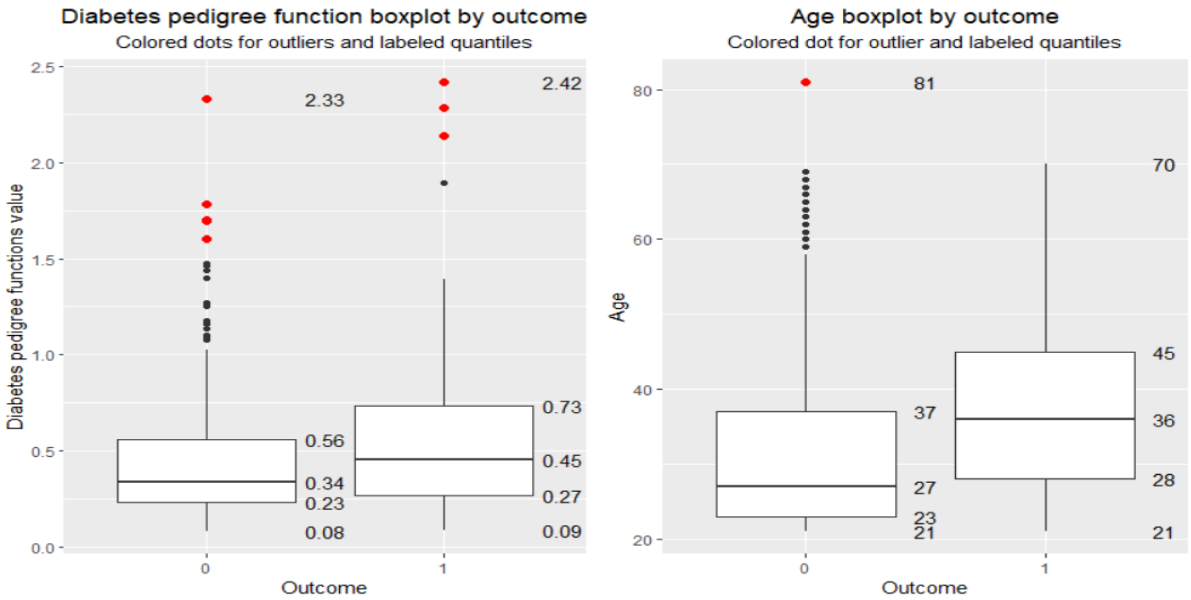
Iš (1 pav.) matome, jog nėštumų skaičius, gliukozės koncentracija kraujyje ir kraujo spaudimas neturi išskirčių tik sąlygines, atitinkamai jų kiekis yra: 15, 17, 15. Insulino koncentraciją žymintis požymis turi 7 išskirtis ir 31 sąlyginę. Galime pastebėti, jog sergančios ir sveikos moterys gerai atsiskiria pagal nėštumų skaičių ir gliukozės koncentracijos atributus, nes sergančių cukriniu diabetu moterų nėštumų skaičiaus mediana yra didesnė 3 vienetais negu sveikų moterų, o gliukozės koncentracijos mediana sergančių moterų atžvilgiu yra 140, sveikų - 107. Sergančio cukriniu diabetu asmens gliuzokės koncentracijos tarpkvartilinis plotas yra platesnis (49,5) nei sveikų (32), t. y. sveikų asmenų gliukozės koncentracijos reikšmės koncentruojasi į

siauresnį intervalą. Šią tendenciją galime pastebėti ir insulino požymyje – sergančių moterų tarpkvartilinis plotas 175 mU / ml, sveikų – 106 mU / ml.




*2 pav. Stačiakampių diagramų panelė, tirianti odos storį ir KMI rodiklius*

Iš sekančios panelės (2 pav.) matome, jog odos storio duomenyse tėra viena sąlyginė išskirtis, o KMI rodiklis turi 1 išskirtį ir 7 sąlygines. Taip pat galime pastebėti, jog pagal odos storio rodiklį grupes atskirti nėra lengva, o pagal KMI rodiklį sergančios ir sveikos moterys vizualiai atsiskiria – pastarosios moterys yra linkusios turėti mažesnį KMI rodiklį. Šios grupės mediana 30,3, o sergančių – 34,3.



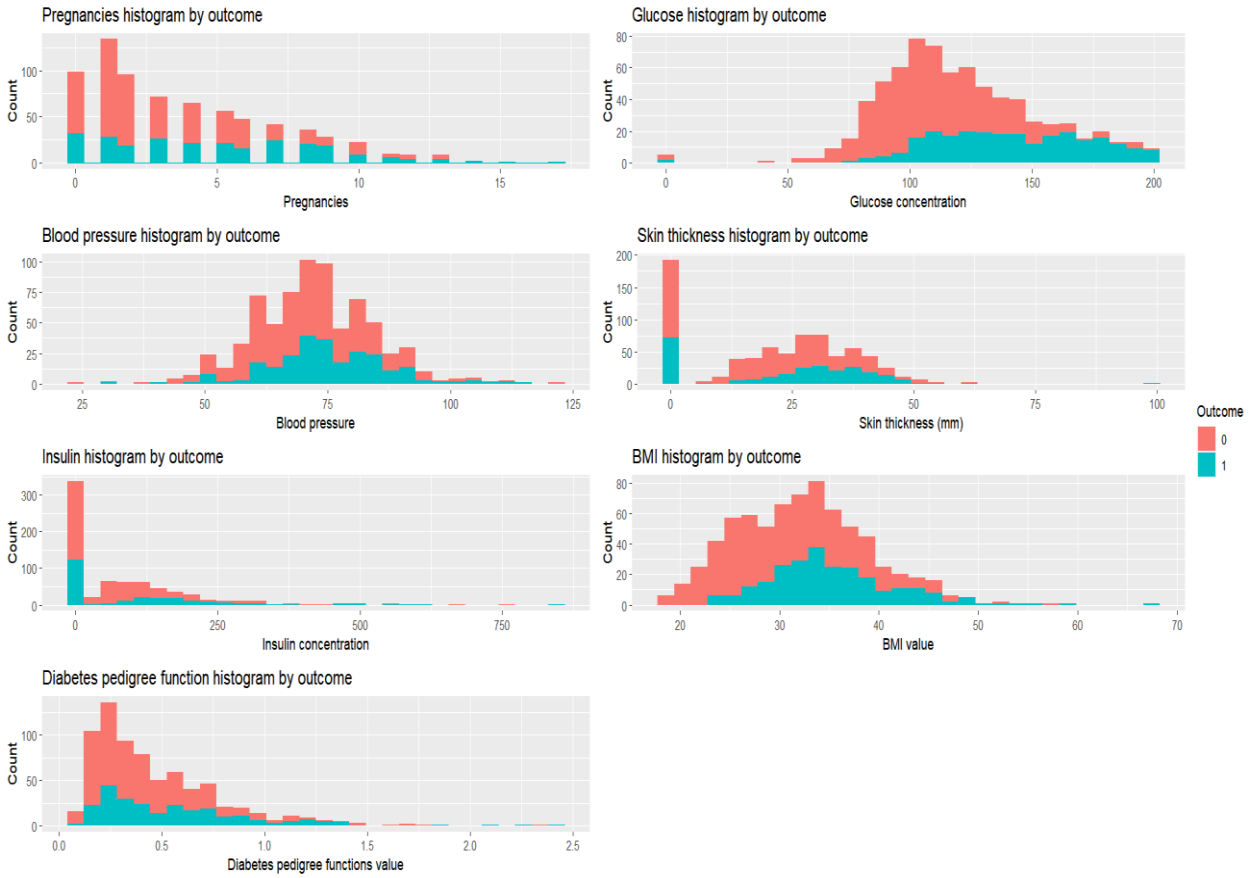

*3 pav. Stačiakampių diagramų panelė, tirianti diabeto atsiradimo funkcijos ir amžiaus požymius*

Iš (3 pav.) matome, jog iš amžiaus požymio lengviau atsiskiria tiriamos grupės, nes sveikų individų mediana 27 metai, o sergančių cukriniu diabetu 36 metai. Galima daryti išvadą, jog vyresnio amžiaus moterys yra labiau linkusios sirgti cukriniu diabetu. Tarp diabeto atsiradimo funkcijos reikšmių yra 8 išskirtys ir 15 sąlyginių, o tarp amžiaus reikšmių yra 1 išskirtis bei 25 sąlyginės. Taip pat galime pastebėti, jog sergančių pacientų diabetinės atsiradimo funkcijos reikšmių tarpkvartilinis plotas 0,46 yra didesnis nei sveikų asmenų 0,33, tai rodo, jog sergančių asmenų funkcijos reikšmės yra mažiau koncentruotos nei sveikų.

Taigi, daugiausiai išskirčių turintys požymiai: diabeto atsiradimo funkcija ir insulino koncentracija, mažiausiai, t. y. nei vienos: nėštumų skaičius, gliukozės koncentracija kraujyje, kraujo spaudimas ir odos storis. Mažiausiai sąlyginių išskirčių turintis požymis yra odos storis, daugiausiai – insulino koncentracija.

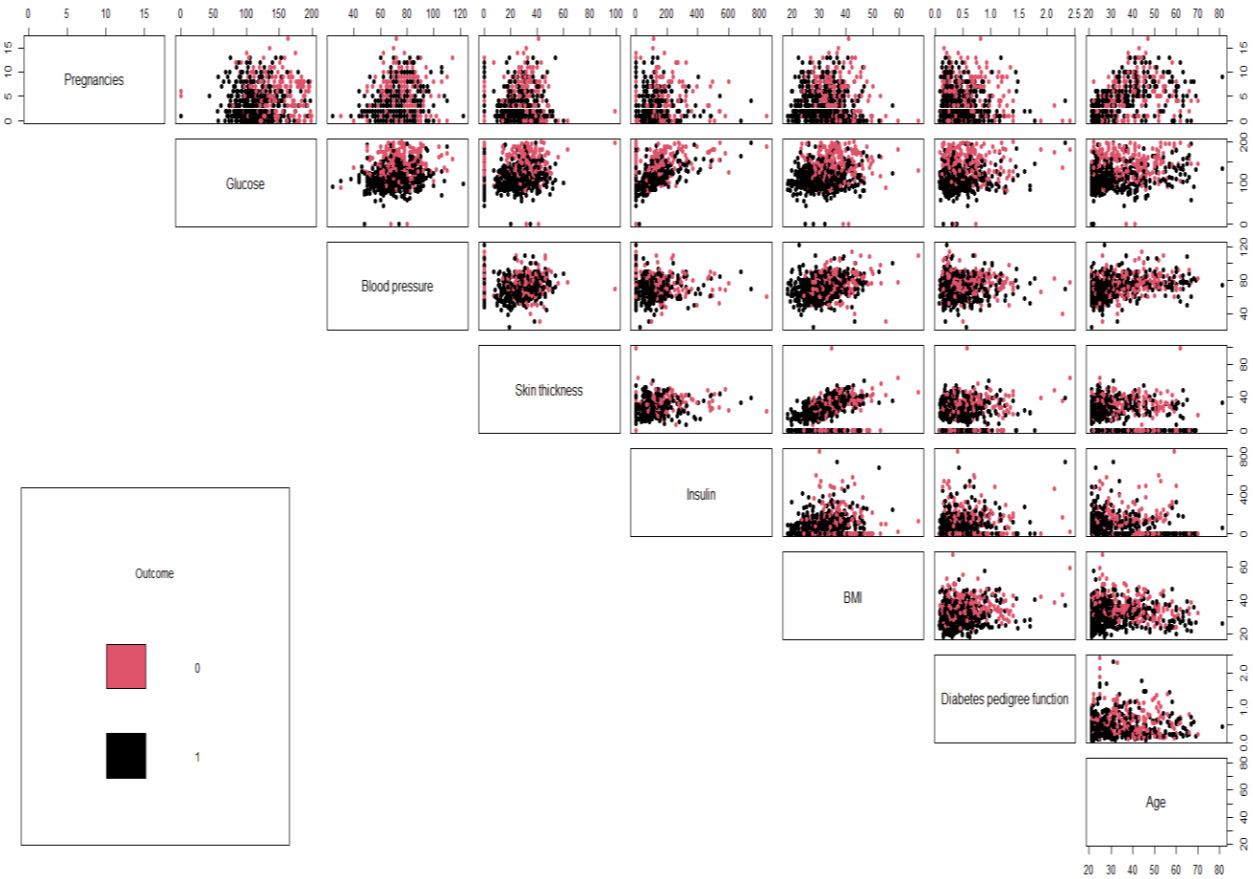
## Papildomi grafikai

Taip pat nusibraižėme keletą papildomų grafikų: histogramas ir sklaidos diagramas, kad galėtumėme detaliau įsigilinti į duomenis.




*4 pav. Histogramų panelė*

Iš (4 pav.) matome, jog sergantys asmenys linkę turėti didesnę gliukozės koncentraciją, KMI rodiklį bei daugiau kartų moterys buvo nėščios. O sveikų moterų yra mažesnė insulino koncentracija bei daugiausiai diabeto atsiradimo funkcijos reikšmės koncentruojasi intervale nuo 0 iki maždaug 0,4.




*5 pav. Visų požymių skalidos diagramos*

Iš (5 pav.) matome, jog visuose požymiuose nėra tikslaus nagrinėjamų grupių atsiskyrimo.

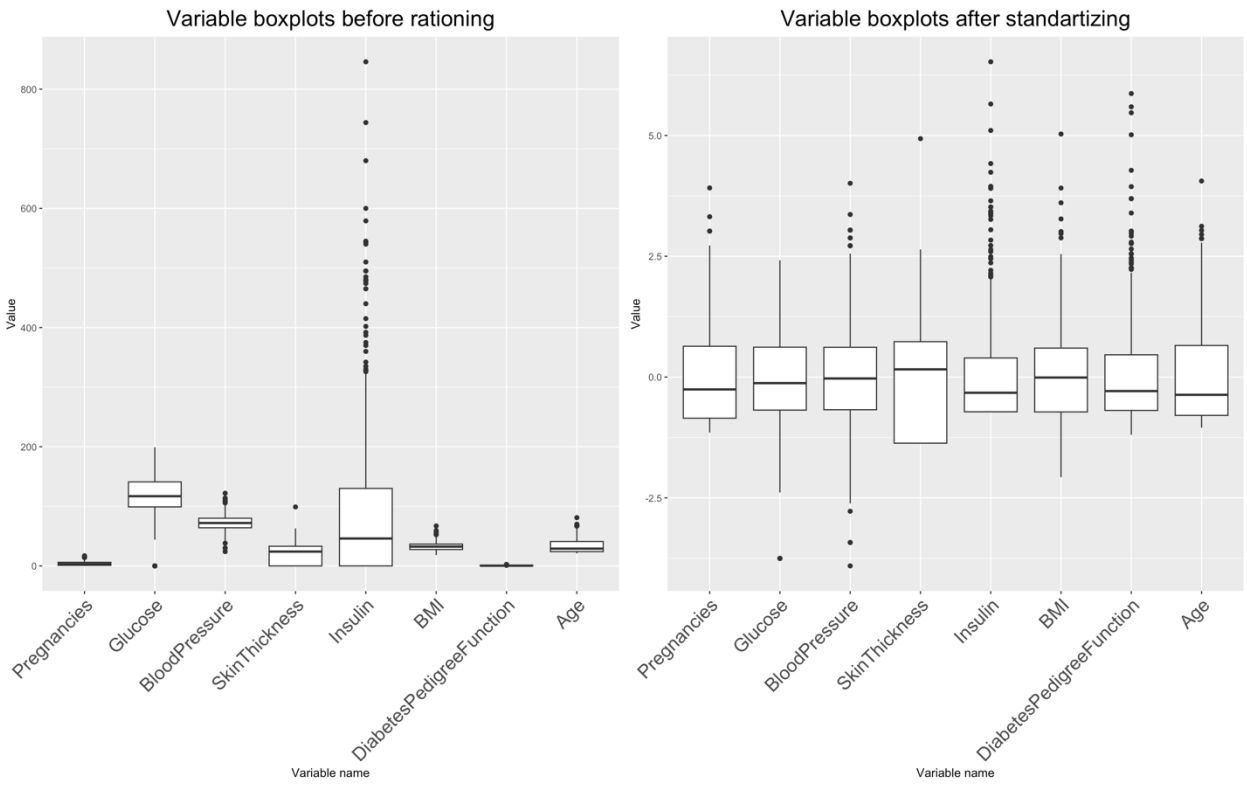
## Duomenų normavimas

Analizuojamos duomenų aibės reikšmės kinta skirtinguose intervaluose, todėl taikysime duomenų normavimą pagal vidurkį ir dispersiją, kuris leidžia suvienodinti reikšmių mąstelius.

Normavimas pagal vidurkį ir dispersiją:

$$x_{norm} = \frac{x - \bar{x}}{\sqrt{\delta^2}},$$

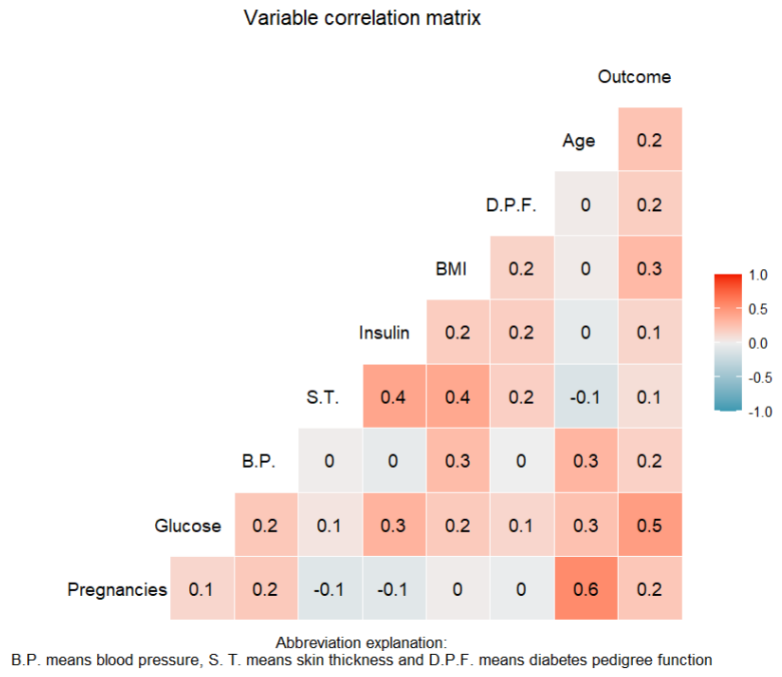
kur $\bar{x}$ – požymio vidurkis, $\delta^2$ - požymio dispersija.




*6 pav. Duomenys prieš normavimą ir po standartizavimo*

Pradinis kiekybinių duomenų aibės požymių pasiskirstymas pavaizduotas stačiakampe diagrama. Pakartotinai pavaizduotas pasiskirstymas atlikus normavimą pagal vidurkį ir disperija. (6 pav.) Matome, kad po standartizavimo labiausiai suvienodėjo kintamųjų vidurkiai.

## Koreliacijos matrica




*7 pav. Koreliacijos matrica*

Tarp skaitinių rodiklių apskaičiuotos Pirsono koreliacijos koeficientų reikšmės. Iš koreliacijos matricos matome, jog stipriausią teigiamą koreliaciją turi nėštumų skaičius ir metai – 0,6 bei gliukozės kiekis su pačiu diagnozavimu, t.y. gliukozės kiekis turi didžiausią įtaką diabeto atsiradimui – 0,5. Neigiama koreliacija yra tarp nėštumų skaičiaus ir odos storio, nėštumų skaičiaus ir insulino kiekio bei odos storio ir metų, jų visų koreliacijos koeficiento reikšmė yra -0,1. Matome, jog stiprių koreliacijų nėra tarp požymių – tai tiesinių išraiškų neturime.

# DIMENSIJŲ MAŽINIMAS

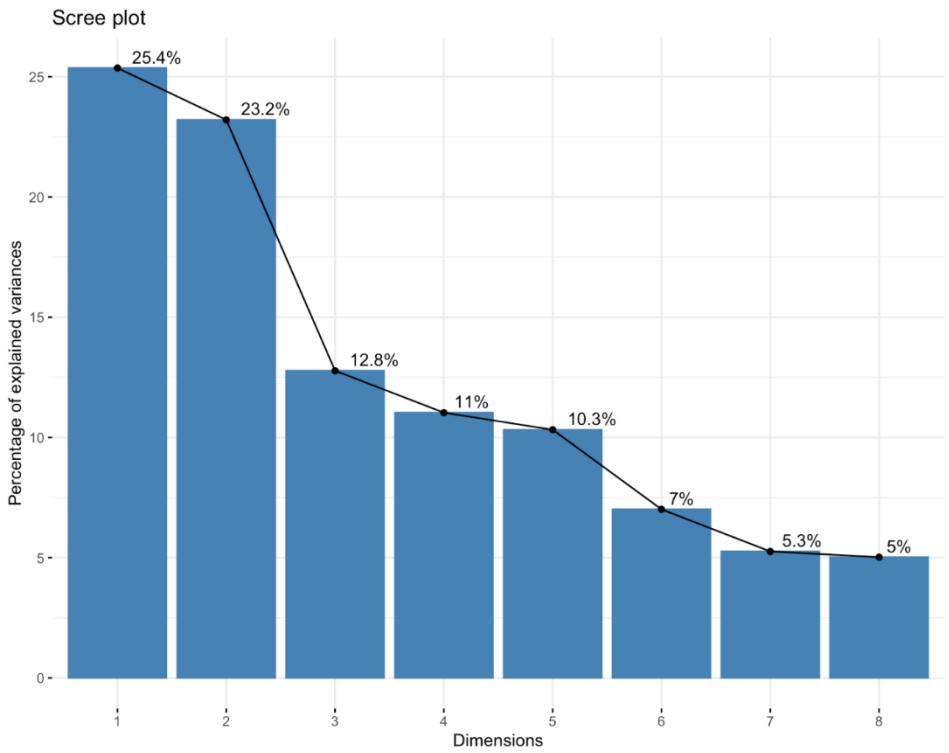
## PCA

Pagrindinių komponenčių analizė (PCA) yra tiesinis dimensijos mažinimo metodas, kuriame duomenų dimensija yra mažinama ieškant tarpusavyje nekoreliuotų ir daugiausiai dispersijos išlaikančių krypčių, vadinamų pagrindinėmis komponentėmis (PC). Kiekviena pagrindinė komponentė yra naujas kintamasis, sudarytas kaip kažkokia pradinių duomenų aibės požymių tiesinė kombinacija (žr. [8]). PCA metodas neturi svarbių parametrų, kuriuos keičiant būtų gaunami skirtingi rezultatai.

Pirmiausia, PCA atlikta su duomenų aibe, kurioje nėra pašalintos išskirtys (pirma pagrindine komponente PC1 paaiškinama 0,25 visos dispersijos, PC2 - 0.23)

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6    PC7     PC8
Standard deviation     1.4245 1.3626 1.0110 0.9395 0.9087 0.74898 0.6487 0.63393
Proportion of Variance 0.2536 0.2321 0.1278 0.1103 0.1032 0.07012 0.0526 0.05023
Cumulative Proportion  0.2536 0.4857 0.6135 0.7238 0.8270 0.89717 0.9498 1.00000
```

*8 pav. Komponenčių svarbumo charakteristikos*




*9 pav. Komponenčių procentinis dispersijos paaiškinimas*

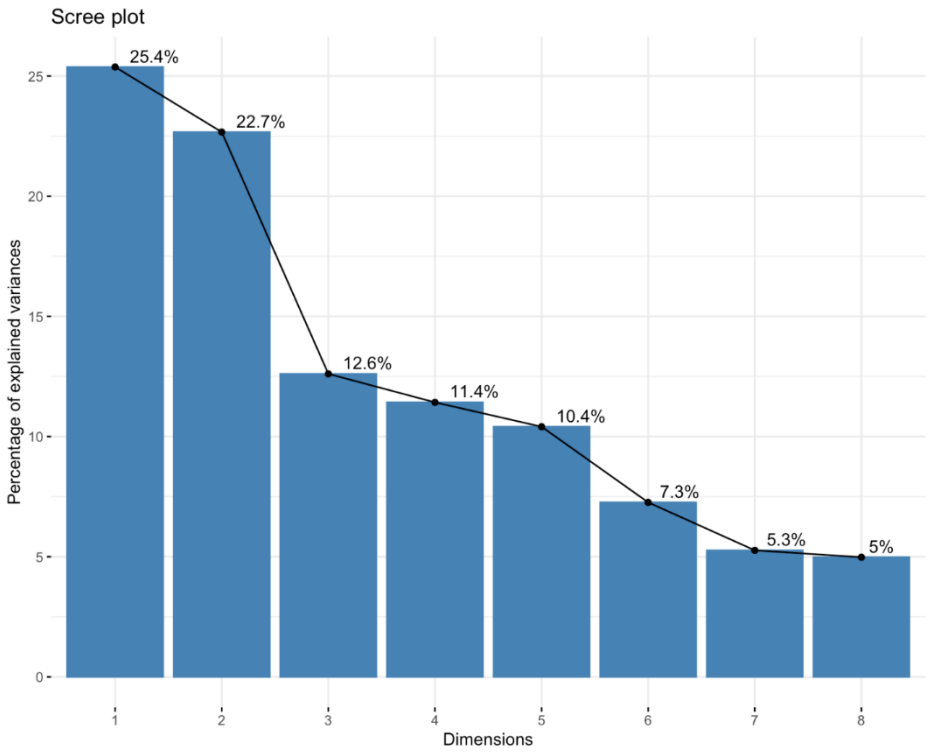
Paliktos ir vizualizuotos pirmos dvi pagrindinės komponentės. Gautuose rezultatuose abi klasės stipriai persidengia.


*10 pav. Pirmųjų komponenčių sklaidos diagrama*

PCA pakartotinai atlikta su duomenų aibe, kurioje yra pašalintos išskirtys (pirma pagrindine komponente PC1 paaiškinama 0,25 visos dispersijos, PCA2 – 0,22).

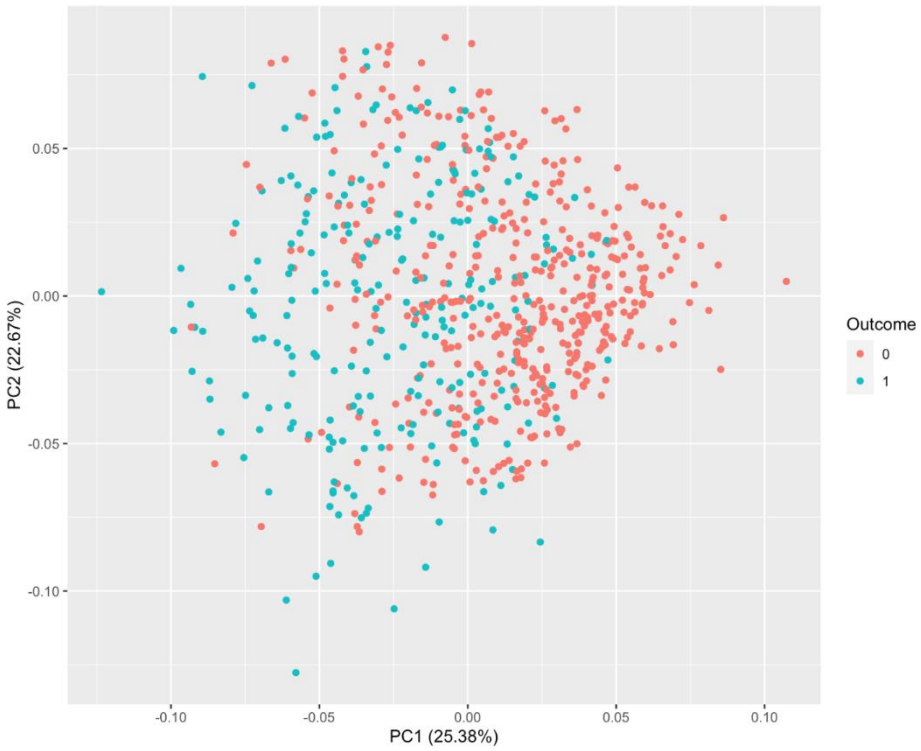
```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.4248 1.3467 1.0044 0.9560 0.9126 0.76219 0.64913 0.63125
Proportion of Variance 0.2538 0.2267 0.1261 0.1143 0.1041 0.07262 0.05267 0.04981
Cumulative Proportion  0.2538 0.4805 0.6066 0.7208 0.8249 0.89752 0.95019 1.00000
```

*11 pav. Komponenčių svarbumo charakteristikos*




*12 pav. Komponenčių procentinis dispersijos paaiškinimas*

Paliktos ir vizualizuotos pirmos dvi pagrindinės komponentės. Gautuose rezultatuose abi klasės taip pat stipriai persidengia.




*13 pav. Pirmųjų komponenčių sklaidos diagrama*

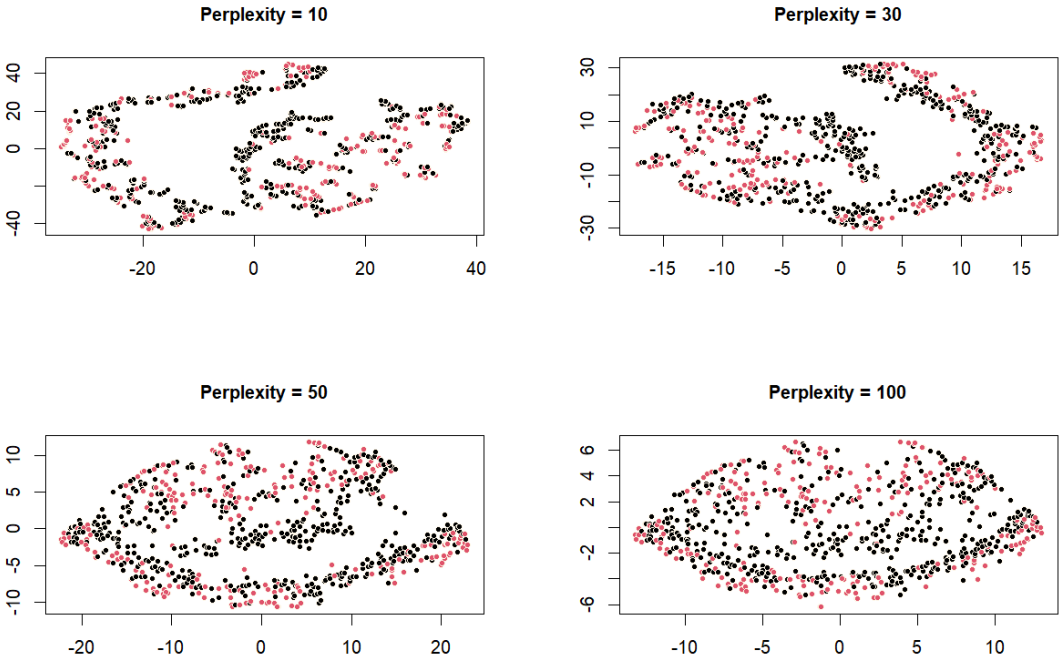
Iš gautų rezultatų matome, jog PCA algoritmas dimensijų mažinimui mūsų duomenims netinka, kadangi neturime tiesinių sąryšių.

## t – SNE

t-SNE ("t-distributed Stochastic Neighbor Embedding") yra netiesinis, neparametrinis dimensijų mažinimo metodas. t-SNE algoritmas pirmiausia pradeda nuo didelės dimensijos Euklidinių atstumų tarp duomenų taškų konvertavimo į sąlygines tikimybes, kurios atspindi panašumus. Tada taškus judina, siekiant išlaikyti mažus atstumus tarp panašių objektų ir didelius atstumus tarp skirtingų objektų. Galiausiai, t-SNE sumažina duomenų rinkinio dimensijas iki reikalingos dimensijos (dažniausiai 2 ar 3 dimensijų) vizualizacijai. t-SNE algoritmas yra naudingas, kai reikia vizualizuoti sudėtingas duomenų struktūras, atskirti jas pagal panašumus ir skirtumus, taip pat identifikuoti ryšius tarp objektų. Tačiau jis turi keletą trūkumų, tokių kaip didelis apskaičiavimo laikas dideliems duomenų rinkiniams ir jautrumas duomenų struktūros pokyčiams. (žr. 22[3], [9])

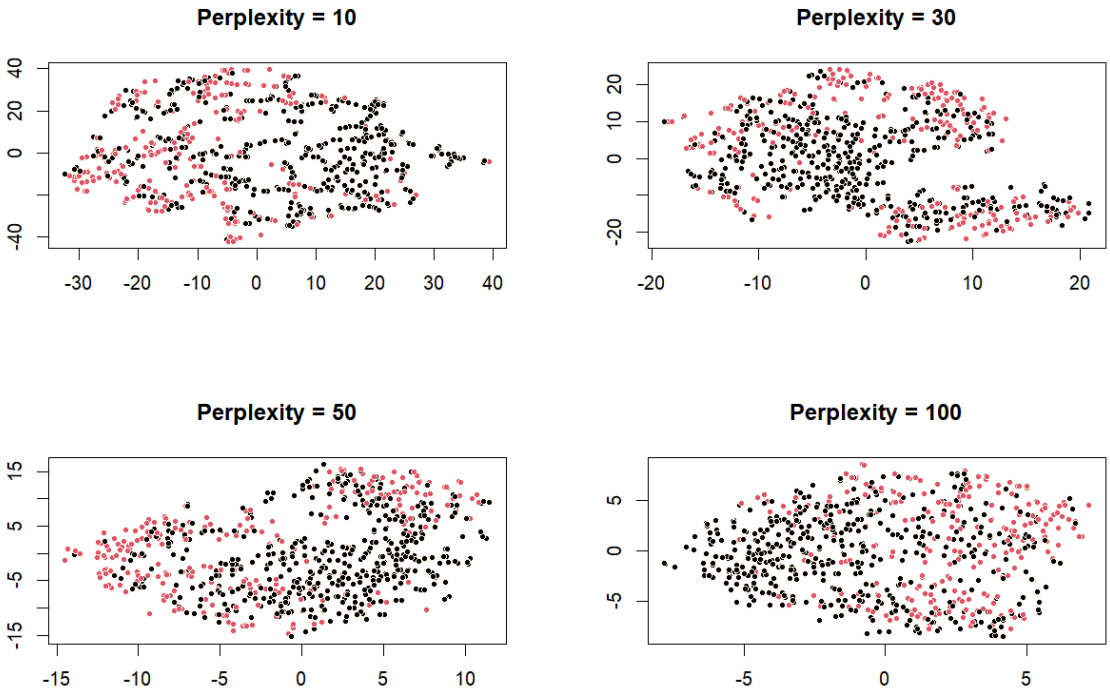
Iš pradžių analizei buvo naudojami pirminiai duomenys, kuriuose nėra pašalintos išskirtys, duomenys ne normuoti ir keičiamas tik perpleksiškumo parametrą, kuris yra vienas svarbiausių

t-SNE parametrų. Jis kontroliuoja Gauso branduolio, naudojamo taškų panašumams apskaičiuoti, plotį ir iš tikrųjų lemia, kiek artimiausių kaimynų pritraukia kiekvienas taškas. (žr. [10])



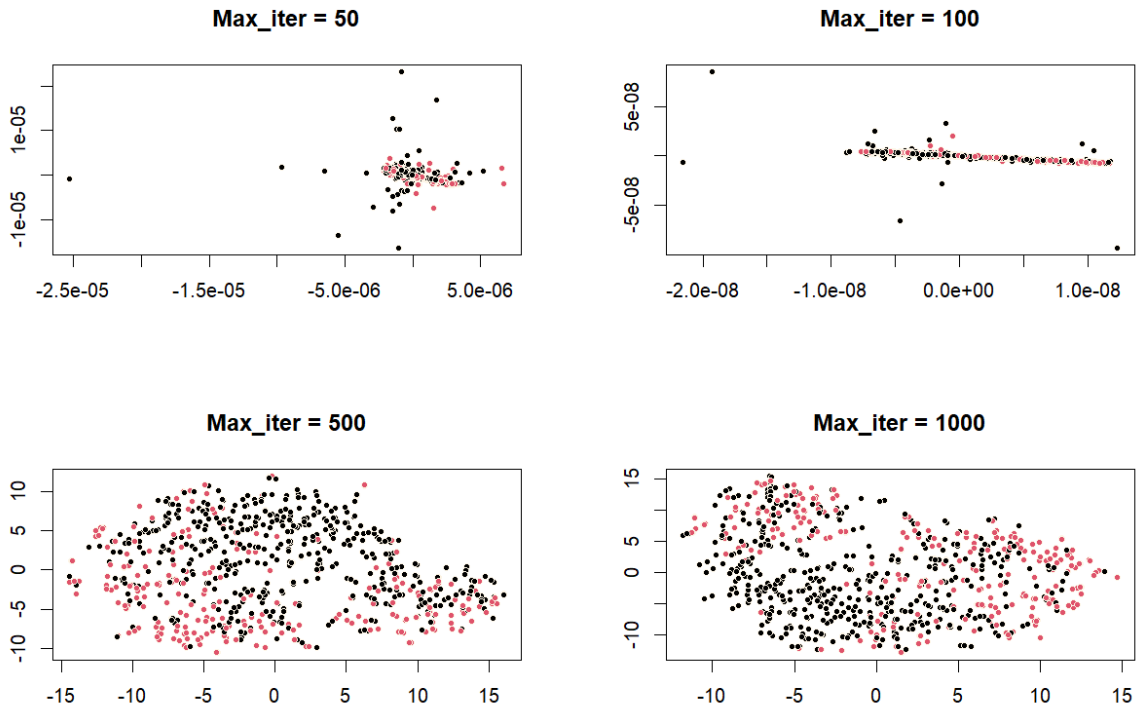

*14 pav. t-SNE grafikai netvarkytiems duomenims*

Galima matyti, kad klasteriai labia persidengia ir sunku juos atskirti. Tie patys parametrai pritaikyti normuotiems duomenims be išskirčių.



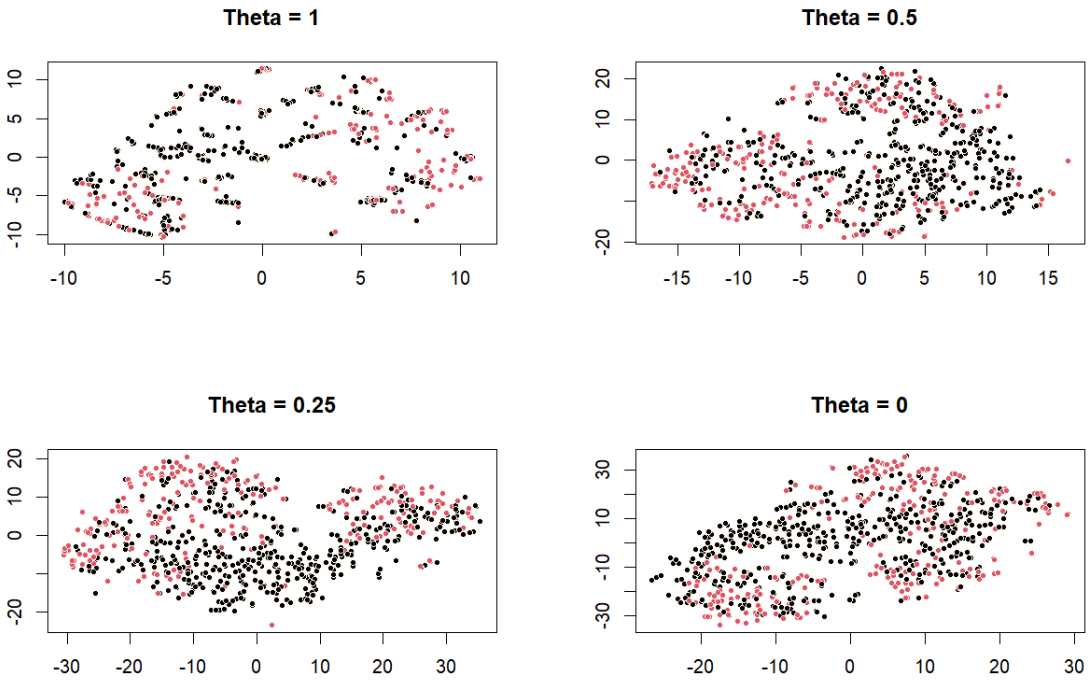

*15 pav. t-SNE grafikai duomenims be išskirčių*

Galima pastebėti, kad klasės minimaliai atiskiria pagal panašumus, tačiau vis dar persidengia. Toliau pasirinkus perpleksiškumą lygų 30, duomenims be išskirčių pritaikytas algoritmas keičiant maksimalų iteracijų skaičių. Duomenys be išskirčių, perplexity = 50, keičiamas maksimalus iteracijų skaičių.




*16 pav. t-SNE grafikai su fiksuotu perplexity parametru*

Klasės geriausiai atsiskiria, kai maksimalus iteracijų skaičius yra 500. Algoritmas pritaikytas duomenims be išskirčių, su nurodytais parametrais – perpleksiškumas lygus 50, maksimalus iteracijų skaičius – 500 ir keičiamas theta parametras – tikslumo rodiklis, kuris kuo didesnis, tuo mažiau tikslus.




*17 pav. t-SNE grafikai su fiksuotais perplexity ir maksimaliu iteracijų skaičiaus parametrais*

Iš rezultatų galima matyti, kad t – SNE minimaliai atskiria klases pagal jų panašumus, tačiau vien šio metodo duomenų interpretacijai nepakanka.

## MDS

Daugiamatė skalė (MDS) yra populiarus metodas, skirtas grafiškai pavaizduoti objektų ryšius daugiamatėje erdvėje. Dimensijų mažinimas taikant MDS pasiekiamas imant pradinį duomenų rinkinį ir apskaičiuojant nepanašumo (atstumo) matą kiekvienai stebėjimų porai palyginti. Tada stebėjimai paprastai grafiškai vaizduojami dvimatėje dimensijoje taip, kad atstumas tarp taškų diagramoje kuo tiksliau atitiktų jų daugiamatį nepanašumą. Šiame darbe naudosime metrikinius ir nemetrikinius MDS algoritmus ir žiūrėsime, kaip keičiant atstumų skaičiavimo metrikas, keičiasi duomenų vizualizavimas. Metrikinėje MDS buvo naudoti atstumų skaičiavimo metodai: kanberos, euklidinis, maksimumo, manhateno bei minkovskio, o nemetrikinėje – kanberos, euklidinis, manhateno, „bray“ ir kulčinskio [6] [12].


*18 pav. Metrikiniai MDS metodai*

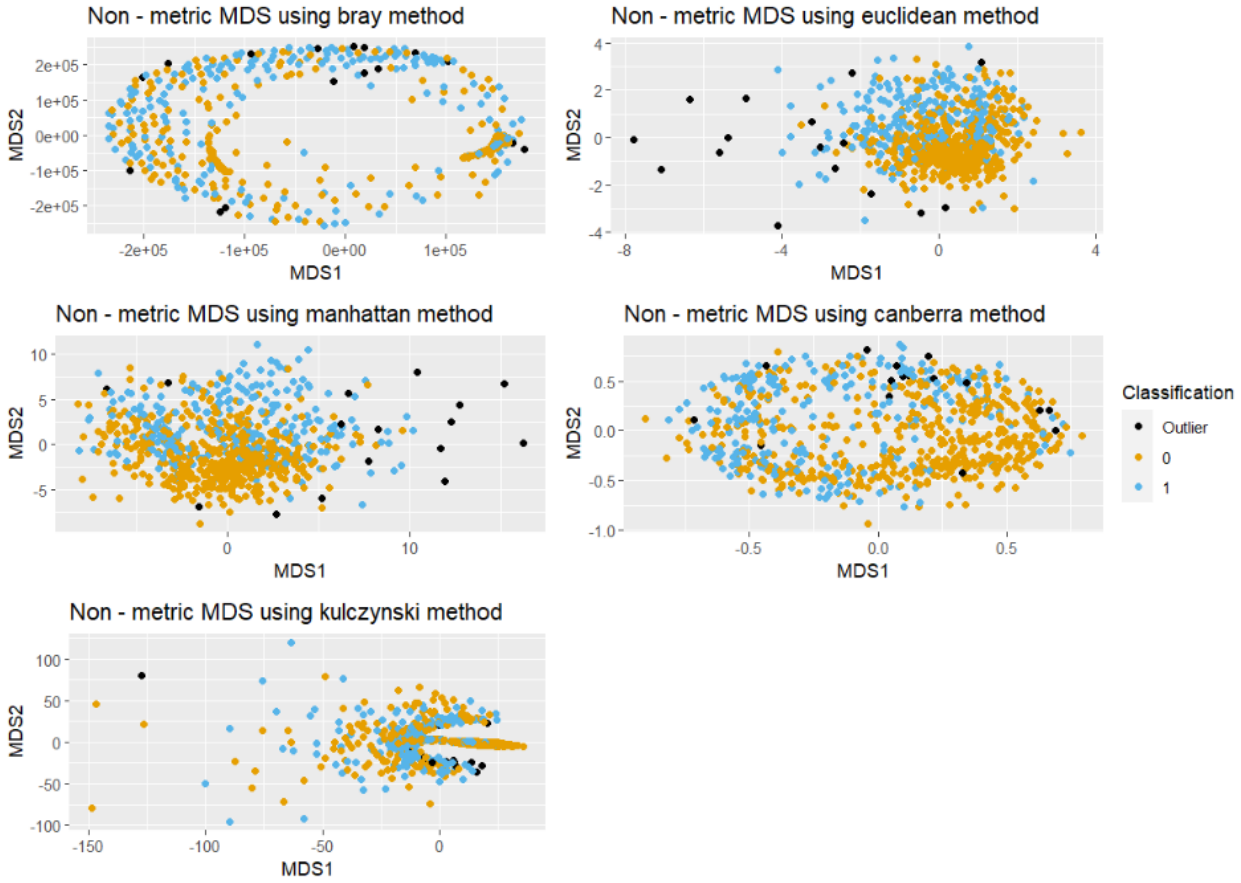
Iš (**Klaida! Nerastas nuorodos šaltinis.**) matome, jog nei su vienu metrikiniu metodu pilnai neatsiskiria grupės. Taip pat galime pastebėti, jog naudojant maksimumo atstumų skaičiavimo metriką geriausiai atsiskiria išsskirtys. Geriausiai grupės išsiskiria taikant euklidinę bei minkovskio atstumų skaičiavimo metrikas. Pastarosios dvi labai panašiai vizualizuoja gautus sumažintos dimensijos duomenis. Blogiausiai vizualizuoja kanberos metodas.

Metrikiniams atstumų radimams suskaičiavome, kiek dispersijos paaiškina paliktos dimensijos. Dimensijų palikome po 3 ir žiūrėjome, kurios dvi geriausiai vizualizuoja duomenis. Kuo didesnis koeficientas tuo geriau, bei kuo tarpusavyje koeficientai panašesni irgi geriau.

*6 lentelė. „GOF" koeficientų reikšmės*

| Atstumų metrika | „GOF" koeficientas | |
|---|---|---|
| Kanbera | 0,190 | 0,263 |
| Euklidinė | 0,486 | 0,486 |
| Maksimumo | 0,140 | 0,214 |
| Manhateno | 0,260 | 0,360 |
| Minkovskio | 0,486 | 0,486 |

Iš (6 lentelė) matome, jog geriausias metodas turėtų būti euklido ir minkovskio atstumų skaičiavimo metrikos. Tą pačią išvadą padarėme iš grafikų.




*19 19 pav. Nemetrikiniai MDS metodai*

Iš (19 19 pav.) matome, jog nei su viena atstumų skaičiavimo metrika grupės visiškai neatsiskiria. Matome, jog išskirtys labiausiai atsiskiria taikant euklidinį bei manheteno atstumų metrikos skaičiavimo būdą. Geriausiai ir aiškiausiai vizualiai grupes atskiria euklidinis metodas, blogiausiai – kanberos ir „bray“ metodai.

Taip pat nemetrikiniams metodams suskaičiavimo streso funkcijos reikšmes, kuo reikšmė mažesnė, tuo geresnis rezultatas [5] [11].

*7 lentelė. Atstumų metrikų streso reikšmės*

| Atstumų metrika | Streso funkcija |
|---|---|
| „Bray“ | 0,259 |
| Euklidinė | 0,210 |
| Manhateno | 0,274 |
| Kanbera | 0,216 |
| Kulčinski | 0,207 |

Iš (7 lentelė) matome, jog didžiausia įtampa gauname naudojant „bray“ metriką, o mažiausią euklidinio ir kulčinski.

# IŠVADOS

Duomenų aibėje nebuvo praleistų reikšmių, tačiau pastebėjome, kad kai kurie įrašai neatitinka logiškos kintamųjų skalės (pvz. kraujo spaudimas ar KMI lygūs 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.

Iš viso buvo rasta 16 išskirčių bei 111 sąlyginių išskirtys. Jos šalintos nebuvo. Daugiausiai išskirčių turintys požymiai: diabeto atsiradimo funkcija ir insulino koncentracija, mažiausiai, t. y. nei vienos: nėštumų skaičius, gliukozės koncentracija kraujyje, kraujo spaudimas ir odos storis. Mažiausiai sąlyginių išskirčių turintis požymis yra odos storis, daugiausiai – insulino koncentracija.

Pradinė analizė parodė, jog padalinus duomenų aibę į grupęs, kai pacientei diagnozuotas diabetas ir kai ne, ir lyginant pagal medianą, visų kovariančių medianinė reikšmė yra didesnė, kai pacientė serga diabetu (turėjusios daugiau nėštumų, gliukozės koncentracija plazmoje didesnė ir t.t.).

PCA atlikus duomenų aibei su išskirtimis (paaiškinta dalis variacijos: PC1 – 0,25, PC2 – 0,23), pašalinus išskirtis rezultatai labai nežymiai suprastėjo (paaiškinta dalis variacijos: PC1 – 0,25, PC2 – 0,22). Rezultatus vizualizavus, pastebimas stiprus klasių persidengimas. Galime teigti, jog naudotas PCA dimensijos mažinimo algoritmas mūsų duomenims netinka, kadangi nėra tiesinių sąryšių, tai parodė ir koreliacijos matrica.

Nors t – SNE metodas yra skirtas netiesiniams duomenims vaizduoti, tačiau gauti rezultatai rodo, kad taikant šį metodą grafikus sunku interpretuoti dėl duomenų persidengimo. Geriausias rezultatas gautas fiksavus perpleksiškumą – 50, maksimalų iteracijų skaičių – 500 ir theta – 0,25.

Naudojant metrikinį MDS dimensijų mažinimo algoritmą geriausiai grupes išskiria euklidinė bei minkovskio atstumų skaičiavimo metrikos, o taikant nemetrikinius MDS dimensijų mažinimo algoritmus aiškiausiai grupes išskiria euklidinis atstumų skaičiavimo metodas. Taip pat galime pastebėti, jog naudojant maksimumo atstumų skaičiavimo metriką geriausiai atsiskiria išskirtys metrikiniame algoritme, o nemetrikiniame geriausiai atsiskiria taikant euklidinį bei

manheteno atstumų metrikos skaičiavimo būdą. Geriausiai ir aiškiausiai vizualiai grupes atskiria euklidinis metodas.

Nagrinėtiems duomenims geriausiai tiko MDS dimensijos mažinimo metodas, kuris geriausiai atskyrė sergančias ir sveikas moteris.

# LITERATŪRA IR ŠALTINIAI


[1] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[2] https://www.datacamp.com/tutorial/pca-analysis-r

[3] https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

[4] https://www.geeksforgeeks.org/how-to-make-pca-plot-with-r/

[5] https://search.r-project.org/CRAN/refmans/AnthropMMD/html/plot.anthropmmd_result.html

[6] https://environmentalcomputing.net/graphics/multivariate-vis/mds/

[7] https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/mds/stress

[8] https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6

[9] https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

[10] https://distill.pub/2016/misread-tsne/

[11] https://stats.stackexchange.com/questions/22019/how-to-calculate-the-r-squared-value-and-assess-the-model-fit-in-multidimensiona

[12] https://rpubs.com/Saskia/520216