VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFOMATIKOS FAKULTETAS

DUOMENŲ MOKSLO BAKALAURAS

DIMENSIJOS MAŽINIMAS KLASIFIKAVIME

Laboratorinis darbas

Atliko: Simona Gelžinytė,

Ugnė Kniukštaitė, Rugilė Bagdonaitė

duomenų mokslas 3 k.

Vilnius, 2023

TURINYS

ĮVADAS

Tikslas

Pritaikyti klasifikavimo algoritmus tiriamai duomenų aibei bei palyginti jų specifikacijas.

Uždaviniai

1. Klasifikavimo šablono sudarymas, klasifikavimo algoritmo tyrimas (Naivojo Bajeso, sprendimų medžio, atsitiktinio miško):
    1.1 Suformuoti klasterizavimui tris duomenų aibes: originalus požymių rinkinys, reikšmingų požymių rinkinys, požymių rinkinys panaudojus dimensijos mažinimo algoritmą;
    1.2 Pateikti tiriamų klasių aprašomąsias statistikas;
    1.3 Padalinti duomenų aibę į mokymo ir testavimo aibes santykiu 80% ir 20%;
    1.4 Apmokyti pasirinktą klasifikatorių ir suklasifikuoti duomenis;
    1.5 Apskaičiuoti sumaišymo matricas;
    1.6 Apskaičiuoti accurary, precision, recall, F1- matus;
    1.7 Įvertinti klasifikavimo rezultatus ir patyrinėti, kaip keičiasi klasifikavimo rezultatai keičiant klasifikatoriaus parametrus;
    1.8 Vizualizuoti klasifikavimo rezultatus;
2. Klasifikavimo kokybės vertinimas. Palyginti skirtingais algoritmais gautus rezultatus ir apibendrinti:
    2.1 Vizualizuoti klasifikavimo rezultatus ROC kreivių grafikais, apskaičiuoti AUC matą;
    2.2 Apibendrinti gautus rezultatus ir pateikti geriausią klasifikavimo rezultatą.

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. Tiriamajame darbe buvo naudojami 3 duomenų rinkiniai: su visais požymiais, reikšmingomis

kovariantėmis (Nėštumas, Gliukozė, KMI, Diabeto susirgimo funkcija) ir sumažintos dimensijos duomenis, naudojant MDS dimensijos mažinimo algoritmą. Suskaičiavę pacientų pasiskirstymą grupėse, gauname, jog duomenų rinkinys yra subalansuotas.

1 lentelė. Pacientų pasiskirstymas grupėse

0	478
1	251

Iš (1 lentelė) galime pasiskaičiuoti, jog mažumos klasės išraiška yra apie 53 %, todėl nesubalansuotumo problemos neturime.

PIRMINIS DUOMENŲ APDOROJIMAS

Tiriamų klasių aprašomosios statistikos

Diagnozuotas diabetas, outcome = 1

2 lentelė. Grupė = 1, aprašomoji statistika

	stand. nuokr.	vidurkis	mediana	min	max
Nėštumas	3,69	5	5	0	17
Gliukozė	32,52	141	140	0	199
Kraujo spaudimas	12,25	75	74	30	114
Odos storis	17,33	24	28	0	60
Insulinas	140,75	107	58	0	846
KMI	6,58	35,35	34,3	22,9	67,1
Diabeto f-ja	0,38	0,56	0,45	0,09	2,42
Amžius	11,08	37	36	21	70

Diabetas nėra diagnozuotas, outcome = 0

3 lentelė. Grupė = 0, aprašomoji statistika

	stand. nuokr.	vidurkis	mediana	min	max
Nėštumas	3,03	3	2	0	13
Gliukozė	26,43	110	107	0	197
Kraujo spaudimas	12,18	71	70	24	122
Odos storis	14,69	20	22	0	60
Insulinas	100,04	72	45	0	744
KMI	6,56	30,96	30,4	18,2	57,3
Diabeto f-ja	0,3	0,43	0,34	0,08	2,33
Amžius	11,55	31	27	21	81

Pasirinkus lyginimo charakteristiką – medianą, matome (2 lentelė - 3 lentelė), 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).

Naudojant visus požymius ir reikšmingas kovariantes, duomenys padalinti į mokymo ir testavimo aibes naudojant santykį 80 : 20. Visų skaitinių požymių matavimo skalės suvienodintos normuojant juos pagal min – max metodą. Normavimui naudotos reikšmės gautos naudojant mokymo aibę.

Norėdamos taikyti klasifikavimo algoritmus sumažintos dimensijos duomenims, iš pradžių susimažinome dimensiją, naudojant MDS algoritmą su Euklidine atstumų skaičiavimo metrika, tada pasidalinimo duomenimis į testavimo ir mokymo aibes santykiu 20 : 80.

KLASIFIKAVIMAS NAUDOJANT NAIVŲJĮ BAJESĄ

Teorinis algoritmo paaiškinimas

Naivojo Bajeso veikimas yra paremtas tikimybėmis:

1. Paėmus konkretų požymių rinkinį yra apskaičiuojama tikimybė, kad objektas priklauso klasei.
2. Tikimybės suskaičiuojamos kiekvienai turimai klasei.
3. Objektas priklauso klasei, kurios tikimybė didžiausia.

Minėtos tikimybės yra apkaičiuojamos remiantis Bajeso teorema, kuri teigia:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)},$$

$$\text{čia } P(A|B) \text{ tikimybė, kad objektas priklausys A klasei, kai jo požymiai yra B,}$$

$$P(A) \text{ tikimybė, kad atsitiktiniu būdu parinktas objektas priklauso A klasei,}$$

$$P(B) \text{ tikimybė, jog atsitiktiniu būdu parinktas objektas turi B požymius}$$

$$(\text{požymių retumo tikimybė}),$$

$$P(B|A) \text{ tikimybė, jog objektas turės B požymius, kai priklauso A klasei.}$$

Jei $P(B|A)$ tikimybė didelė, tai leidžia užtikrinčiau tikėti, jog nagrinėjamas objektas priklauso A klasei. Priešingu atveju – implikuoja, jog mūsų objektas grečiausiai nebus iš A klasės.

Naivusis Bajesas naudoja prielaidą, jog tarpusavyje požymiai yra nepriklausomi. [1][3]

Algoritmo veikimo pavyzdys

Tarkime turime dvi gyvūnų klases: kates ir šunis. Kiekvienas objektas turi šiuos požymius: svoris (lengvas, vidutinis, sunkus) ir vikrumo lygį (pasyvus, vidutinis, judrus). Norime sužinoti, kaip klasifikuoti naują objektą, kurio parametrai būtų: vidutinio svorio nejudrus gyvūnas. Užsirašome tikimybes, kad gyvūnas priklausys kiekvienai galimai klasei:

$$P(\text{Katė}|\text{Vidutinis svoris, nejudrus}) = \frac{P(\text{Vidutinis svoris, nejudrus}|\text{Katė}) \times P(\text{Katė})}{P(\text{Vidutinis svoris, nejudrus})},$$

$$P(\text{Šuo}|Vidutinis\ svoris, nejudrus) = \frac{P(Vidutinis\ svoris, nejudrus|\text{Šuo}) \times P(\text{Šuo})}{P(Vidutinis\ svoris, nejudrus)}.$$

Mes ieškosime, kuri tikimybė yra didesnė. Tarkime norime sužinoti, ar naujas gyvūnas labiau tikėtina, kad yra katė negu šuo, t. y. pasinaudodami Bajeso terorema mes tikrinsime:

$$\frac{P(Vidutinis\ svoris, nejudrus|Katė) \times P(Katė)}{P(Vidutinis\ svoris, nejudrus)} > \frac{P(Vidutinis\ svoris, nejudrus|\text{Šuo}) \times P(\text{Šuo})}{P(Vidutinis\ svoris, nejudrus)}.$$

Kadangi tikimybė visada yra teigiamas dydis galime pasinaikinti vardiklį ir gausime paprastesnę formą:

$$P(Vidutinis\ svoris, nejudrus|Katė) \times P(Katė) > P(Vidutinis\ svoris, nejudrus|\text{Šuo}) \times P(\text{Šuo}).$$

Iš tikimybių teorijos yra žinoma, jog galioja tokia formulė:

$$P(A, B) = P(A) \times P(B|A)\ arba\ P(A, B) = P(B) \times P(A|B).$$

Pritaikius šią formulę mūsų uždaviniui, gauname:

$$P(Vidutinis\ svoris, nejudrus|Katė) \times P(Katė) = P(Vidutinis\ svoris, nejudrus, katė),$$

$$P(Vidutinis\ svoris, nejudrus|\text{Šuo}) \times P(\text{Šuo}) = P(Vidutinis\ svoris, nejudrus, \text{šuo}).$$

Taip pat iš tikimybių teorijos yra žinoma, jog galioja tokia formulė:

$$P(A, B, C) = P(A|B, C) \times P(B|C) \times P(C).$$

Pasinaudoję šia formule bei, kad požymiai nepriklausomi ($P(A|B) = P(A)$) dėl Bajeso prielaidos, gauname:

$$P(Vidutinis\ svoris, nejudrus, katė) = P(Vidutinis\ svoris|Katė) \times P(Nejudrus|Katė) \times P(Katė),$$

$$P(Vidutinis\ svoris, nejudrus, \text{šuo}) = P(Vidutinis\ svoris|\text{Šuo}) \times P(Nejudrus|\text{Šuo}) \times P(\text{Šuo}).$$

Skaičiuoti tokias tikimybes ir palyginti žymiai lengviau nei pradines.

Tarkime iš imties žinome, jog $P(Vidutinis\ svoris|Katė) = 0{,}3, P(Nejudrus|Katė) = 0{,}6, P(Katė) = 0{,}5, P\big(Vidutinis\ svoris\big|Šuo\big) = 0{,}5, P\big(Nejudrus\big|Šuo\big) = 0{,}1, P\big(Šuo\big) = 0{,}5.$

Gauname, jog:

$$P(Katė|Vidutinis\ svoris, nejudrus) = 0{,}3 \times 0{,}6 \times 0{,}5 \ = 0{,}09,$$

$$P\big(Šuo|Vidutinis\ svoris, nejudrus\big) = 0{,}5 \times 0{,}1 \times 0{,}5 \ = 0{,}025.$$

Su šiomis tikimybėmis labiau tikėtina, jog naujas gyvūnas yra katė. [3]

Galimi Naiviojo Bajeso tipai

Yra išskiriami 5 Naiviojo Bajeso tipai:

- Gauso – tinkamas, kai požymiai yra tolydieji bei pasiskirstę pagal normulųjį dėsnį.
- Daugialypis – dažniausiai naudojamas žodžių klasifikacijai.
- Bernulio – naudingas, kai požymiai turi po 2 reikšmes.
- Pilnasis – rekomenduojamas, kai duomenų rinkinys yra nesubalansuotas.
- Kategorinis – tinkamas, kai nepriklausomi kintamieji yra kategoriniai. [2][4]

Algoritmo privalumai ir trūkumai

Privalumai:

- Tinka nedidelėms duomenų imtims.
- Greičiau konverguoja nei kiti modeliai.
- Tinkamas kategoriniams ir tolydiesiems duomenims.
- Nėra jautrus nereikšmingoms kovariantėms.
- Tinkamas tekstų analizei.

Trūkumai:

- Remiasi požymių nepriklausomumu, o tai dažnai yra neišpildyta sąlyga.
- Ilgai dirba su duomenimis, kuriuose yra daug požymių.

- „0 dažnumo“ problema, jei mokymosi duomenyse nebuvo konkrečios priklausomo kintamojo reikšmės, o testavimo duomenyse ji buvo. Šios reikšmės įgijimo tikimybė yra 0. Šiai problemai spręsti yra naudojamos pataisos.[1][2]

Gauso Naviojo Bajeso pritaikymas duomenims

Naudojant šį algoritmą nevykdėme parametrų paieškos, neradome hiperparametrų, kuriuos būtų galima reguliuoti. Buvo gauti identiški rezultatai taikant Gauso Naivųjį Bajeso algoritmą visiem požymiams ir reikšmingiems.

4 lentelė. Maišos matrica gauta naudojant Gauso Naivųjį Bajeso algoritmą visiems požymiams ir reikšmingiems

		Prognozuoti		
		1	**0**	
Tikri	**1**	50	0	**Jautrumas**: 1,00
	0	96	0	**Specifiškumas**: 0,00
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: NaN	**Bendras tikslumas**: 0,34

Iš (4 lentelė) matome, jog nebuvo pasiekta gerų rezultatų – bendras tikslumas tėra 0,34. Modelis visus testavimo duomenis priskyrė prie sergančių, visiškai neatpažino sveikų asmenų. F_1 buvo gautas 0,51, kuris yra apskaičiuojamas:

$$F_\beta = \frac{(1 + \beta^2) \times (tikslumas\ (angl. precision) \times jautrumas)}{\beta^2 \times tikslumas\ (angl. precision) + jautrumas}, \text{čia } \beta \text{ gali būti } 0{,}5, 1, 2.$$

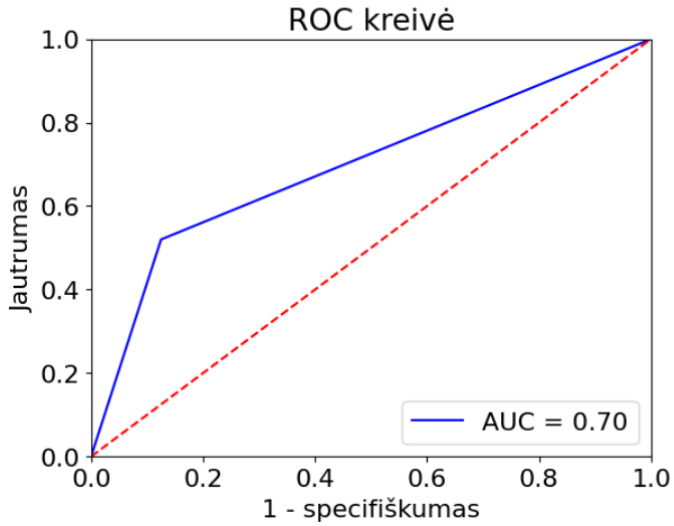
ROC kreivė šiem pritaikytiems modeliams nebuvo informatyvi, nes plotas po ja buvo tik 0,5.

Klasifikuojant sumažintos dimensijos duomenis, naudojant MDS algoritmą su Euklidine atstumų metrika, buvo gauti geresni rezultatai.

5 lentelė. Maišos lentelė sumažintos dimensijos duomenims, naudojant Gauso Naivųjį Bajeso metodą

		Prognozuoti		
		1	**0**	
Tikri	**1**	26	24	**Jautrumas**: 0,52
	0	12	84	**Specifiškumas**: 0,88
		Tikslumas (precision): 0,68	**Neigiama prognostinė vertė**: 0,78	**Bendras tikslumas**: 0,75

Iš (5 lentelė) matome, jog testiniai duomenis buvo žymiai geriau suklasifikuoti lyginant su didesnios dimensijos duomenimis (bendras tikslumas 0,75 lyginant su 0,34). Galime pastebėti, jog geriausiai sekėsi atpažinti nesergančius asmenis (specifiškumas 0,88), o sergančius klasifikatorius sunkiau atskyrė (jautrumas tik 0,52). F_1 buvo gautas 0,59.



1 pav. ROC kreivė sumažintos dimensijos duomenims, naudojant Gauso Naivųjį Bajeso algoritmą

Iš (1 pav.) matome, jog pasiektas neblogas plotas po kreive – 0,7.

6 lentelė. Gautų Gauso klasifikatorių palyginimas

Naudoti požymiai	Bendras tikslumas	Jautrumas	Tikslumas (angl. *precision*)	F_1	AUC
8	0,34	1,00	0,34	0,51	0,50
4	0,34	1,00	0,34	0,51	0,50
2	0,75	0,52	0,68	0,59	0,70

Iš (6 lentelė) matome, jog geriausiai Gauso Bajeso klasifikatorius pasirodė sumažintos dimensijos duomenims. Galime teigti, jog duomenis geriau klasifikuoja mažesnėje dimensijoje.

Daugialypio Naviojo Bajeso pritaikymas duomenims

Šiam klasifikavimo algoritmui radome 3 reguliuojamus parametrus:

- *Alpha* – šiam parametrui ėmėme reikšmes iš intervalo [0,1].
- *Force_alpha* – šio parametro galimos reikšmės taip arba ne. Jei *alpha* parametro reikšmė bus mažesnė nei 1×10^{-10} ir *Force_alpha* bus ne, tai bus palikta 1×10^{-10} reikšmė.
- *Fit_prior* - šio parametro galimos reikšmės taip arba ne. Jei reikšmė ne, tikimybė priklausyti klasei bus imama iš tolygiojo skirstinio.

Geriausias parametrų rinkinys buvo ieškomas naudojantis parametrų gardele ir taikant 5 – kryžminę validacijos patikrą. Renkant parametrus buvo gauti geriausi rezultatai su keliais skirtingais parametrų rinkiniais. Sumažintos dimensijos duomenims neišėjo pritaikyti daugialypio Naiviojo Bajeso metodo, nes klasifikavimo algoritmas nepriėmė neigiamų reikšmių, kurios atsirado po dimensijos mažinimo.

7 lentelė. Keli rezultatai iš parametrų paieškos, naudojant daugialypį Naivųjį Bajeso metodą

Naudojant visus požymius				
Alpha	*Force_alpha*	*Fit_prior*	Tikslumas validavimo aibėje	Rinkinio gerumo vieta
0	Taip	Taip	0,655	1
0,1	Taip	Taip	0,655	1
0,1	Ne	Ne	0,607	21
0,2	Ne	Taip	0,609	15
0,25	Taip	Taip	0,655	1
0,25	Ne	Taip	0,609	15
0,3	Taip	Ne	0,655	1
Naudojant reikšmingas kovariantes				
Alpha	*Force_alpha*	*Fit_prior*	Tikslumas validavimo aibėje	Rinkinio gerumo vieta
0	Taip	Taip	0,655	1
0	Ne	Taip	0,521	35
0,1	Taip	Ne	0,655	1
0,1	Ne	Taip	0,521	35
0,3	Taip	Ne	0,655	1
0,3	Ne	Taip	0,521	35

0,6	Ne	Taip	0,526	27
0,7	Taip	Ne	0,655	1

Nors matome, jog su keliais parametrų rinkiniais yra gaunami geriausi galimi rezultatai iš (7 lentelė), pasilikome toliau dirbti su rinkiniu: $\{Alpha: 0, Force\_alpha: taip, Fit\_prior: taip\}$ abiejuose duomenų rinkiniuose.

8 lentelė. Maišos lentelė, naudojant visus požymius ir daugialypį Naiviojo Bajeso klasifikatorių

		Prognozuoti		
		1	**0**	
Tikri	**1**	24	26	**Jautrumas**: 0,48
	0	47	49	**Specifiškumas**: 0,51
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0,65	**Bendras tikslumas**: 0,50

Iš (8 lentelė) matome, jog klasifikatorius prognozuoja negerai – bendras tikslumas tėra 0,5, dažnai klysta – iš visų prognozuotų sergančių iš tikrųjų sirgo tik 34 proc.. F_1 buvo gautas 0,40, plotas po ROC kreive taip pat buvo mažas tik 0,50.

9 lentelė. Maišos lentelė, naudojant reikšmingus požymius ir daugialypį Naiviojo Bajeso klasifikatorių

		Prognozuoti		
		1	**0**	
Tikri	**1**	0	50	**Jautrumas**: 0,00
	0	0	96	**Specifiškumas**: 0,51
		Tikslumas (precision): NaN	**Neigiama prognostinė vertė**: 0,66	**Bendras tikslumas**: 0,66

Iš (9 lentelė) matome, jog, nors ir klasifikatoriaus bendras tikslumas padidėjo, naudojant tik reikšmingus požymius ir lyginant su rezultatais, kai naudojami visi požymiai (0,66 ir 0,50), tačiau algoritmas visiškai neatpažįsta sergančių asmenų. F_1 balo nebuvo galima apskaičiuoti, nes tikslumo (angl. Precision) taip pat nebuvo galima suskaičiuoti bei jautrumas buvo 0,00. Plotas po ROC kreive taip pat buvo mažas tik 0,50.

10 lentelė. Daugialypio Naiviojo Bajeso klasikatorių palyginimas

Naudoti požymiai	**Bendras tikslumas**	**Jautrumas**	**Tikslumas (angl. precision)**	F_1	**AUC**
8	0,50	0,48	0,34	0,40	0,50
4	0,66	0,00	NaN	NaN	0,50

Iš (10 lentelė) negalime, kad teigti, jog kažkuris klasifikatorius geresnis – nei vienas neparodė gerų rezultatų.

Pilnojo Naviojo Bajeso pritaikymas duomenims

Šiam klasifikavimo algoritmui radome 2 reguliuojamus parametrus:

- *Alpha* – šiam parametrui ėmėme reikšmes iš interval $[0,1]$.
- *Force_alpha* – šio parametro galimos reikšmės taip arba ne. Jei *alpha* parametro reikšmė bus mažesnė nei 1×10^{-10} ir *Force_alpha* bus ne, bus palikta 1×10^{-10} reikšmė.

Geriausias parametrų rinkinys buvo ieškomas naudojantis parametrų gardele ir taikant 5 – kryžminę validacijos patikrą. Renkant parametrus buvo gauti keli geriausi rezultatai su skirtingais parametrų rinkiniais. Sumažintos dimenijos duomenims neišėjo pritaikyti pilnojo Naiviojo Bajeso metodo, nes klasifikavimo algoritmas nepriėmė neigiamų reikšmių, kurios atsirado po dimensijos mažinimo.

11 lentelė. Keli rezultatai iš parametrų paieškos, naudojant pilnąjį Naivųjį Bajeso metodą

Naudojant visus požymius			
Alpha	*Force_alpha*	Tikslumas validavimo aibėje	Rinkinio gerumo vieta
0,00	Taip	0,607	23
0,20	Ne	0,609	17
0,30	Taip	0,609	17
0,40	Ne	0,610	15
0,45	Taip	0,612	1
0,50	Ne	0,612	1
0,55	Taip	0,612	1
Naudojant reikšmingus požymius			
Alpha	*Force_alpha*	Tikslumas validavimo aibėje	Rinkinio gerumo vieta

0	Ne	0,521	15
0,3	Ne	0,521	15
0,6	Taip	0,526	5
0,9	Ne	0,528	3
1	Taip	0,530	1
1	Ne	0,530	1

Nors matome, jog su keliais parametrų rinkiniais yra gaunami geriausi galimi rezultatai iš (11 lentelė), pasilikome toliau dirbti su rinkiniais:

- $\{Alpha: 0{,}45,\ Force\_alpha: taip\}$ naudojant visus požymius.
- $\{Alpha: 1{,}00,\ Force\_alpha: taip\}$ naudojant reikšmingus požymius.

12 lentelė. Maišos lentelė, naudojant visus požymius ir pilnąjį Naiviojo Bajeso klasifikatorių

		Prognozuoti		
		1	**0**	
Tikri	**1**	24	26	**Jautrumas**: 0,48
	0	49	47	**Specifiškumas**: 0,49
		Tikslumas (precision): 0,33	**Neigiama prognostinė vertė**: 0,64	**Bendras tikslumas**: 0,49

Iš (12 lentelė) matome, jog klasifikatorius prognozuoja netiksliai – bendras tikslumas tėra 0,49, dažnai klysta – iš visų prognozuotų sergančių iš tikrųjų sirgo tik 33 proc.. F_1 buvo gautas 0,39, plotas po ROC kreive taip pat buvo mažas tik 0,48.

13 lentelė. Maišos lentelė, naudojant reikšmingus požymius ir pilnąjį Naiviojo Bajeso klasifikatorių

		Prognozuoti		
		1	**0**	
Tikri	**1**	0	50	**Jautrumas**: 0,00
	0	0	96	**Specifiškumas**: 0,51
		Tikslumas (precision): NaN	**Neigiama prognostinė vertė**: 0,66	**Bendras tikslumas**: 0,66

Iš (13 lentelė) matome, jog, nors ir klasifikatoriaus bendras tikslumas padidėjo, naudojant tik reikšmingus požymius ir lyginant su rezultatais, kai naudojami visi požymiai (0,66

ir 0,50), tačiau algoritmas visiškai neatpažįsta sergančių asmenų. F_1 balo nebuvo galima apskaičiuoti, nes tikslumo (angl. *precision*) taip pat nebuvo galima suskaičiuoti bei jautrumas buvo 0,00. Plotas po ROC kreive taip pat buvo mažas tik 0,50. Gautas lygiai toks pat rezultatas kaip su daugialypiu Naiviuoju Bajesu reikšmingoms kovariantėms.

14 lentelė. Pilnojo Naiviojo Bajeso klasikatorių palyginimas

Naudoti požymiai	**Bendras tikslumas**	**Jautrumas**	**Tikslumas (angl. precision)**	F_1	**AUC**
8	0,49	0,48	0,33	0,39	0,48
4	0,66	0,00	NaN	NaN	0,50

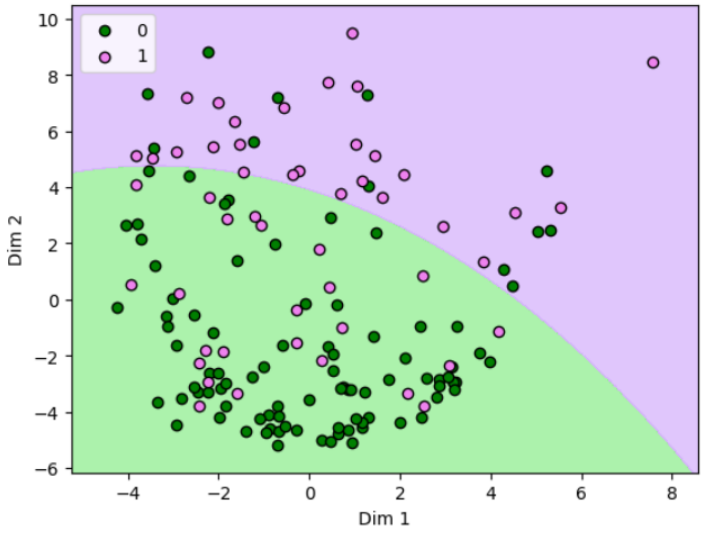
Iš (14 lentelė) negalime, kad teigti, jog kažkuris klasifikatorius geresnis – nei vienas neparodė gerų rezultatų.

Geriausias rezultatas

15 lentelė. Bajeso klasifikatorių palyginimas

	Naiviojo Bajeso rūšis	**Bendras tikslumas**	F_1	**AUC**
Visi požymiai	Gauso	0,34	0,51	0,50
	Daugialypis	0,50	0,40	0,50
	Pilnasis	0,49	0,43	0,5
Reikšmingos kovariantės	Gauso	0,34	0,51	0,50
	Daugialypis	0,66	NaN	0,50
	Pilnasis	0,66	NaN	0,50
Sumažintos dimensijos	Gauso	**0,75**	**0,59**	**0,70**

Iš (15 lentelė) matome, jog geriausias pasiektas rezultatas buvo naudojant sumažintos dimensijos duomenis ir Gauso Naivųjį Bajeso klasifikatorių, kurio bendras tikslumas buvo 0,75, jautrumas 0,52, tikslumas (angl. *precision*) 0,68, F_1 0,59 ir AUC 0,70.



2 pav. Klasifikavimo rezultatai testinei duomenų aibei

Iš (2 pav.) matome, jog, nors ir pasiekti geri klasifikavimo rodikliai, tačiau algoritmas tiksliai nesugeba nupiešti skiriamojo paviršiaus.

12 stebėjimų buvo klaidingai priskirta diabeto liga, o 24 pacientams klaidingai liga nenustatyta testinėje aibėje.

16 lentelė. Klaidingų 1 aprašomosios statistikos palyginimas su esamais 1

Klaidingų 1					
	stand. nuokrypis	vidurkis	mediana	min	max
Nėštumas	5,02	6	5	0	13
Gliukozė	35,90	136	142	82	189
Kraujo spaudimas	15,17	85	85	52	110
Odos storis	18,07	27	27	0	60
Insulinas	106,96	74	0	0	310
KMI	10,25	36,52	35,80	22,20	57
Diabeto f-ja	0,51	0,72	0,64	0,16	1,70
Amžius	14,55	39	39	22	60
Tikrų 1					
	stand. nuokrypis	vidurkis	mediana	min	max
Nėštumas	3,69	5	5	0	17

Gliukozė	32,52	141	140	0	199
Kraujo spaudimas	12,25	75	74	30	114
Odos storis	17,33	24	28	0	60
Insulinas	140,75	107	58	0	846
KMI	6,58	35,35	34,3	22,9	67,1
Diabeto f-ja	0,38	0,56	0,45	0,09	2,42
Amžius	11,08	37	36	21	70

Iš (16 lentelė. Klaidingų 1 aprašomosios statistikos palyginimas su esamais 1) galime matyti, jog beveik visiems požymiams, išskyrus insulinui, klaidingų 1 medianos reikšmės buvo panašios, o kai kur net didesnės už tikrų 1.

17 lentelė. Klaidingų 0 aprašomosios statistikos palyginimas su esamais 0

Klaidingų 0					
	stand. nuokrypis	vidurkis	mediana	min	max
Nėštumas	2,44	4	4	0	8
Gliukozė	20,56	140	140	106	178
Kraujo spaudimas	10,42	69	70	48	86
Odos storis	15,20	18	21	0	45
Insulinas	79,70	68	0	0	210
KMI	5,49	32,63	31,95	23,80	44,00
Diabeto f-ja	0,15	0,46	0,43	0,26	0,76
Amžius	10,68	33	29	22	60
Tikrų 0					
	stand. nuokrypis	vidurkis	mediana	min	max
Nėštumas	3,03	3	2	0	13
Gliukozė	26,43	110	107	0	197
Kraujo spaudimas	12,18	71	70	24	122
Odos storis	14,69	20	22	0	60
Insulinas	100,04	72	45	0	744
KMI	6,56	30,96	30,4	18,2	57,3
Diabeto f-ja	0,3	0,43	0,34	0,08	2,33
Amžius	11,55	31	27	21	81

Iš (17 lentelė. Klaidingų 0 aprašomosios statistikos palyginimas su esamais 0) matome, jog amžiui, kraujo spaudimui, diabeto funkcijai, odos storiui medianų reikšmės buvo panašios klaidingų ir tikrų 0, o kitų požymių medianos reikšmės buvo didesnės nei tikrų 0.

KLASIFIKAVIMAS NAUDOJANT SPRENDIMŲ MEDĮ

Sprendimų medžio metodas sukuria klasifikavimo modelius medžio struktūros pavidalu. Duomenų rinkinys sprendimo mazguose yra suskaidomas į vis mažesnius poaibius, kartu palaipsniui kuriant susijusį sprendimų medį. Galutinis rezultatas yra medis su sprendimo mazgais ir lapų mazgais. Lapo mazgo, į kurį pateko klasifikuojamas objektas, reikšmė atitinką modelio priimtą sprendimą. Sprendimo mazgai yra konstruojami pasirenkant tokį kintamąjį, pagal kurio reikšmes galima geriausiai padalinti duomenų rinkinį.

Sprendimų medžiai lengvai suprantami ir interpretuojami. Kitas iš šio metodo privalumų yra beveik nereikalingas pradinis duomenų apdorojimas: pateikiami duomenys neturi būti vienodoje skalėje, priklausomai nuo metodo implementacijos gali būti pateikiami objektai su praleistomis požymių reikšmėmis, nebūtina perkoduoti kategorinių kintamųjų, savaime atliekamas daugiau negu dviejų klasių klasifikavimas.

Sprendimų medžiai sugeba lengvai prisitaikyti prie struktūrų, esančių mokymo duomenyse, tačiau rezultatai itin stipriai priklauso nuo to, kokie duomenys buvo mokymo aibėje. Dėl šios priežasties sprendimų medžiu tikėtina gauti prastesnius rezultatus klasifikuojant prieš tai nematytus stebėjimus. Su šia problema susijęs modelio parametrų parinkimas, pavyzdžiui: *max_depth* kontroliuoja maksimalų medžio gylį, *min_samples_split* parametru parenkamas minimalus stebėjimų kiekis, reikalingas norint dar kartą skaidyti duomenų aibę, *min_samples_leaf* – minimalus reikiamas stebėjimų skaičius medžio lapuose. [5][7]

Kadangi sprendimų medžiai požymį naudoja konstruoti sprendimų mazgui tik jeigu jis gerai atskiria klases (sprendimų medžiai atlieka savaiminį požymių atrinkimą), todėl nesitikima gauti rezultatų pagerėjimo atrenkant požymių poaibį. Optimalių parametrų ieškota naudojant parametrų tinklelį *max_depth* = {4,5,6}, *min_samples_split* = {5,10,15}, *max_features* = {0.4,0.6,0.8}.

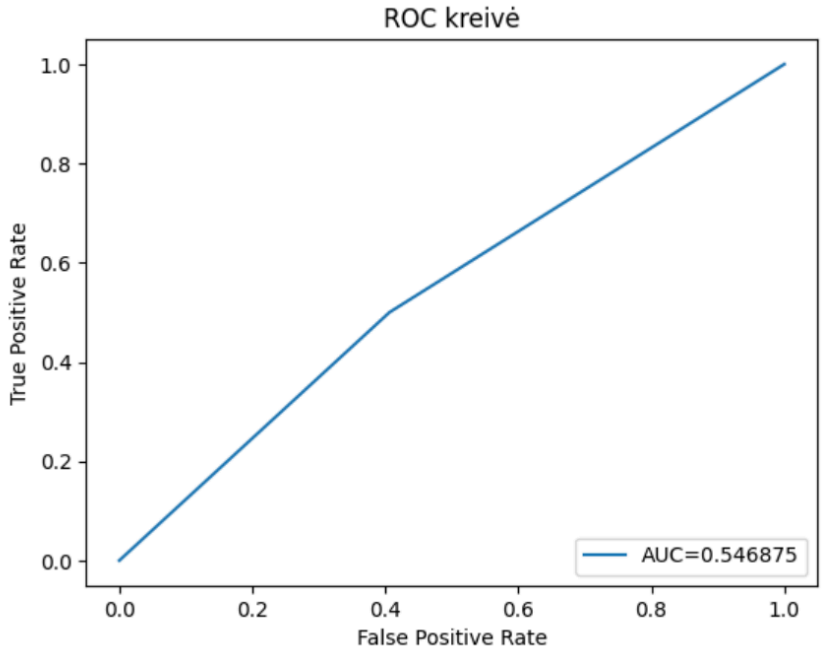
Visam duomenų rinkiniui

Iš pradžių algoritmas buvo pritaikytas normuotai duomenų aibei pagal nutylėjimą nurodytais parametrais. Gauta klasifikavimo matrica, kurios bendras tikslumas – 0,56, o F_1 įvertis – 0,44. Nubraižius ROC kreivė, gautas 0,55 AUC įvertis.

18 lentelė. Maišos matrica visam duomenų rinkiniui

		Prognozuoti		
		1	**0**	
Tikri	**1**	25	25	**Jautrumas**: 0,50
	0	39	57	**Specifiškumas**: 0,59
		Tikslumas (precision): 0,39	**Neigiama prognostinė vertė**: 0,69	**Bendras tikslumas**: 0,56

Iš (18 lentelė) matome, jog pasiektas bendras tikslumas nėra didelis 0,56, taip pat klasifikatorius neatskiria sergančių pacientų testavimo aibėje – jautrumas tik 0,5. Gautas F_1 balas: 0,44.



3 pav. ROC kreivė visam duomenų rinkiniui

Taip pat iš (3 pav.) galime pastebėti, jog ROC kreivė yra tik minimaliai išlenkta ir plotas po ja tik 0,55.

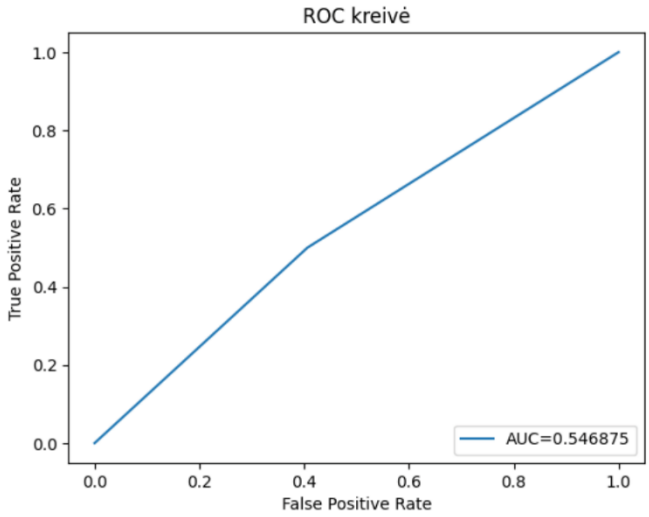
Toliau buvo pritaikyta *GridsearchCV* funkcija geriausiam hiperparametrų rinkiniui rasti, naudojant 5 – kryžminę validaciją. Geriausias rezultatas su hiperparametrais:

```
{
    'max_depth': 4,
    'max_features': 0.6,
    'min_samples_split': 5
}
```

Su šiais hiperparametrais gautas klasifikavimo matricos bendras tikslumas – 0,66 (žr. 19 lentelė), o F_1 įvertis – 0. Gautas AUC įvertis – 0,55 (žr. 4 pav.). Bendras tikslumas geresnis nei su numatytaisiais parametrais, tačiau F_1 įvertis prastas.

19 lentelė. Maišos matrica su geriausiais hiperparametrais

		Prognozuoti		
		1	**0**	
Tikri	**1**	0	50	**Jautrumas**: 0
	0	0	96	**Specifiškumas**: 1
		Tikslumas (precision): NaN	**Neigiama prognostinė vertė**: 0,66	**Bendras tikslumas**: 0,66



4 pav. ROC kreivė su geriausiais hiperparametrais

5 geriausios klasifikavimo hiperparametrų kombinacijos:

20 lentelė. Geriausios klasifikavimo hiperparametrų kombinacijos

Max_depth	***Max_features***	***Min_samples_split***	**Tikslumas**
4	0,6	5	0,66
4	0,6	10	0,66
4	0,6	15	0,66
4	0,8	15	0,64
6	0,6	15	0,55

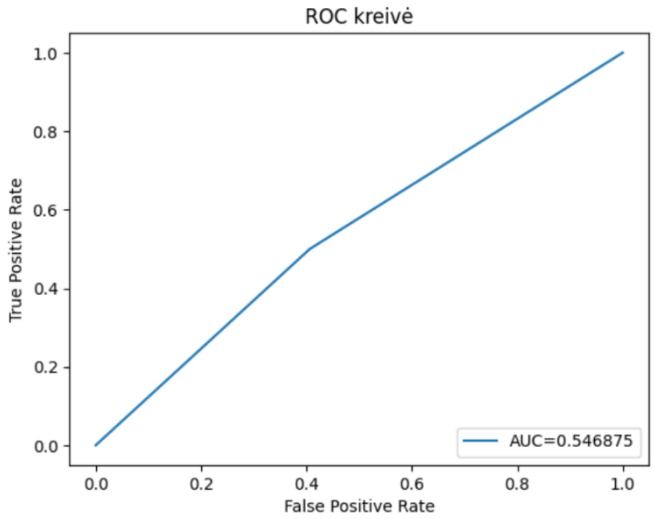
Galime pastebėti iš (20 lentelė), jog *Min_samples_split* neturi įtakos klasifikavimo tikslumui, jam keičiantis tikslumas visada išlieka toks pats.

Reikšmingos kovariantės

Toliau klasifikavimas atliktas atrinktoms reikšmingoms kovariantėms – gliukozės kiekiui, neštumų skaičiui, KMI ir diabeto susirgimo funkcijai. Su numatytaisiais parametrais klasifikavimo matricos bendras tikslumas – 0,34, F_1 įvertis – 0,5, AUC – 0,55 (žr. 5 pav.). Galima pastebėti, kad visi stebėjimai yra priskirti vienai klasei (žr.).

21 lentelė. Maišos matrica reikšmingoms kovariantėms

		Prognozuoti		
		1	**0**	
Tikri	**1**	50	0	**Jautrumas**: 1,00
	0	96	0	**Specifiškumas**: 0,00
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: NaN	**Bendras tikslumas**: 0,34



5 pav. ROC kreivė reikšmingoms kovariantėms

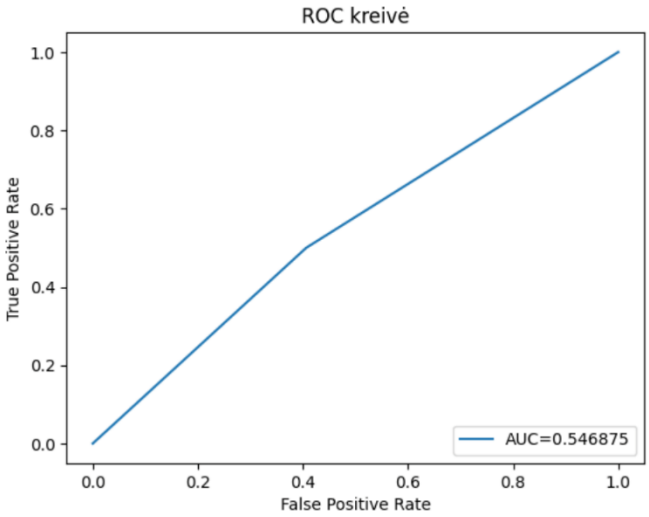
Geriausias rezultatas su parametrais:

```
{
    'max_depth': 5,
    'max_features': 0.4,
    'min_samples_split': 5
}
```

Didžioji dauguma stebėjimų vėl buvo priskirti vienai klasei, bendras klasifikavimo tikslumas yra šiek tiek geresnis (0,37) (žr. 22 lentelė. Maišos matrica su geriausiais hiperparametrais), F_1 yra 0,49, o AUC išlieka toks pats – 0,55 (žr. 6 pav.).

22 lentelė. Maišos matrica su geriausiais hiperparametrais

		Prognozuoti		
		1	**0**	
Tikri	**1**	45	5	**Jautrumas**: 0,90
	0	87	9	**Specifiškumas**: 0,10
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0,64	**Bendras tikslumas**: 0,37



6 pav. ROC kreivė su geriausiais hiperparametrais

5 geriausios klasifikavimo hiperparametrų kombinacijos:

23 lentelė. Geriausios klasifikavimo hiperparametrų kombinacijos

Max_depth	***Max_features***	***Min_samples_split***	**Tikslumas**
5	0,4	5	0,37
5	0,4	10	0,37
4	0,4	15	0,34
4	0,4	10	0,34
4	0,4	15	0,34

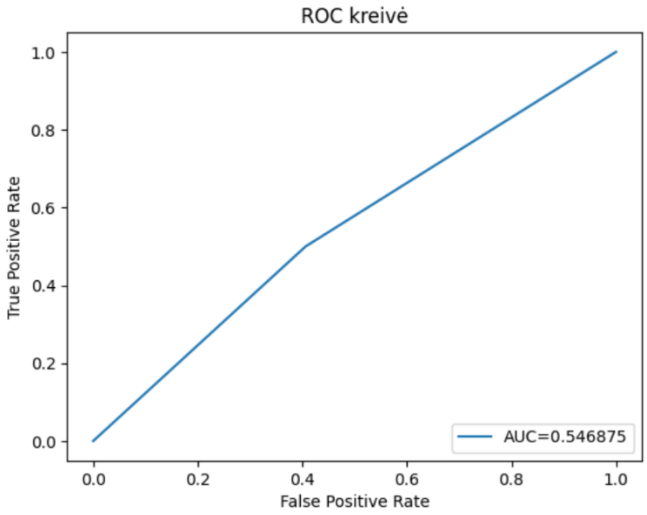
Vėlgi galime pastebėti iš (23 lentelė), jog *Min_samples_split* neturi įtakos klasifikavimo tikslumui, *Max_features* taip pat, o pasirinkus *Max_depth* = 5, gauname geriausią tikslumą.

Sumažintos dimensijos

Duomenų aibė buvo sumažinta iki dviejų dimensijų pritaikius MDS euklidinį metodą ir jam buvo pritaikytas klasifikavimas. Su numatytaisiais parametrais gautas iki šiol geriausias bendras tikslumas – 0,63 (žr. 24 lentelė), F_1 – 0,48 ir AUC įvertis – 0,55 (žr. 7 pav.).

24 lentelė. Maišos matrica sumažintos dimensijos

		Prognozuoti		
		1	**0**	
Tikri	**1**	25	25	**Jautrumas**: 0,50
	0	29	67	**Specifiškumas**: 0,70
		Tikslumas (precision): 0,46	**Neigiama prognostinė vertė**: 0,73	**Bendras tikslumas**: 0,63



7 pav. ROC kreivė sumažintai dimensijai

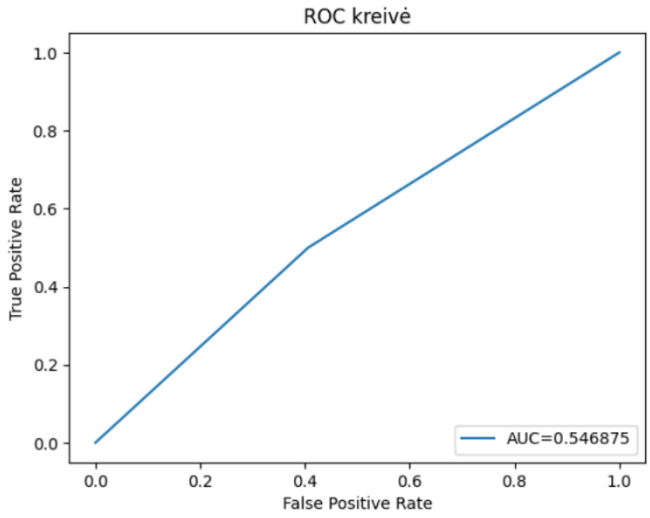
Geriausias rezultatas su parametrais:

```
{
    'max_depth': 4,
    'max_features': 0.4,
    'min_samples_split': 5
}
```

Daugiau nei puse stebėjimų buvo priskirti teisingai klasei. Gautas geriausias tikslumas - 0,68 (žr. 25 lentelė), F_1 – 0,5, AUC nepakito – 0,55 (žr. 8 pav.).

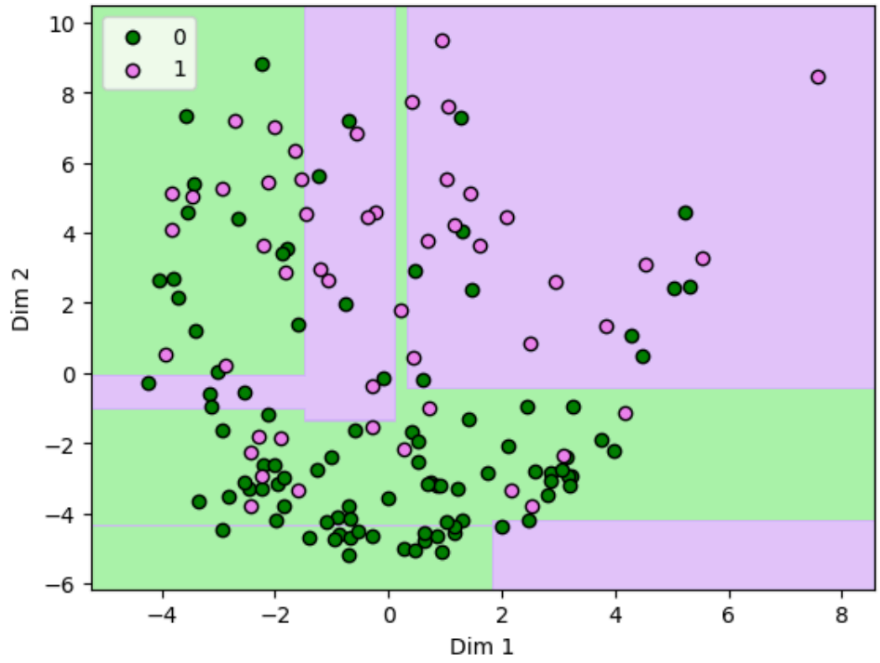
25 lentelė. Maišos matrica su geriausiais hiperparametrais

			Prognozuoti	
		1	**0**	
Tikri	**1**	23	27	**Jautrumas**: 0,46
	0	19	77	**Specifiškumas**: 0,8
		Tikslumas (precision): 0,55	**Neigiama prognostinė vertė**: 0,74	**Bendras tikslumas**: 0,68



8 pav. ROC kreivė sumažintai dimensijai

Toliau pateiktas sumažintos dimensijos su *GridSearchCV* atrinktais geriausiais hiperparametrais gautas klasifikavimas. Iš (9 pav.) galima matyti, kad dauguma taškų yra atitinkamos spalvos fone, reiškiančios, kad taškas buvo klasifikuotas teisingai.



9 pav. Klasifikavimo tikslumas sklaidos diagramoje

5 geriausios klasifikavimo hiperparametrų kombinacijos:

26 lentelė. Geriausios klasifikavimo hiperparametrų kombinacijos sumažintai dimensijai

Max_depth	***Max_features***	***Min_samples_split***	**Tikslumas**
4	0,4	5	0,68
4	0,4	10	0,68
4	0,6	5	0,68
4	0,6	10	0,68
4	0,8	5	0,68

Iš (27 lentelė) atsispindi visi sprendimų medžio metodu gauti klasifikavimo rezultatai. Visiems duomenų rinkiniams AUC įvertis gavosi toks pats. Tiksliausiai buvo klasifikuojamas sumažintos dimensijos duomenų rinkinys, F_1 įvertis taip pat gautas geriausias.

27 lentelė. Klasifikavimo rezultatų palyginimas

	Visi požymiai		**Reikšmingos kovariantės**		**Sumažintos dimensijos**	
	Numatyti parametrai	*Grid SearchCV*	Numatyti parametrai	*Grid SearchCV*	Numatyti parametrai	*Grid SearchCV*
Bendras tikslumas	0,56	0,66	0,34	0,37	0,63	0,68
F_1	0,44	0	0,5	0,49	0,48	0,5
AUC	0,55	0,55	0,55	0,55	0,55	0,55

KLASIFIKAVIMAS NAUDOJANT ATSITIKTINĮ MIŠKĄ

Atsitiktinis miškas yra vienas populiariausių klasifikavimo algoritmų. Jis sprendžia tiek klasifikavimo, tiek regresijos uždavinius. Šis algoritmas yra prižiūrimo mokymosi ir sukuria mišką su daug medžių, kuo didesnis medžių skaičius miške, tuo tikslesni rezultatai yra gaunami. Šis algoritmas nėra jautrus trūkstamoms reikšmėms. *„RandomForest“* turi beveik tokius pačius parametrus, kaip sprendimų medis ar pakavimo (angl. *Bagging*) klasifikatorius. *„RandomForest“* prideda papildomo atsitiktinumo modeliui auginant medžius. Užuot ieškojęs svarbiausio atributo, algoritmas skaido duomenis ir ieško geriausių savybių tarp atsitiktinių funkcijų pogrupio. Tai lemia didelę įvairovę, kuri paprastai lemia geresnį modelį. Todėl, *„RandomForest“* algoritmas atsižvelgia tik į atsitiktinį funkcijų pogrupį. [9] [10]

Atsitiktinio miško algoritmai turi tris pagrindinius hiperparametrus, kuriuos reikia nustatyti prieš pradedant mokymą. Tai - mazgų dydis, medžių skaičius ir atrinktų požymių skaičius. Toliau atsitiktinio miško klasifikatorius gali būti naudojamas regresijos arba klasifikavimo uždaviniams spręsti.

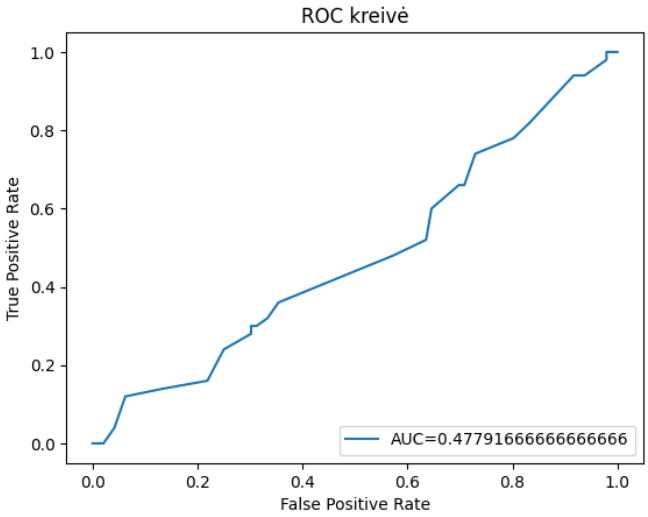
Atsitiktinio miško algoritmą sudaro sprendimų medžių rinkinys, o kiekvieną medį sudaro duomenų imtis, paimta iš mokymo aibės. Iš šios mokymo imties trečdalis yra atidedama kaip testo duomenys, vadinamoji išvestinė imtis (angl. *out-of-bag, oob*). Tada, taikant požymių paketą, į duomenų aibę įvedamas dar vienas atsitiktinumo atvejis, taip padidinant įvairovę ir sumažinant sprendimų medžių koreliaciją. Priklausomai nuo problemos tipo, prognozės nustatymas skirsis. Atliekant regresijos užduotį, atskiriems sprendimų medžiams bus išvedamas vidurkis, o atliekant klasifikavimo užduotį prognozuojama klasė bus nustatoma pagal daugumos balsus, t. y. pagal dažniausiai pasitaikantį kategorinį kintamąjį. Galiausiai išvestinė imtis naudojama kryžminiam tikrinimui, galutinai patvirtinant tą prognozę. [8]

Visam duomenų rinkiniui

Iš pradžių algoritmas buvo pritaikytas normuotai duomenų aibei. Gauta klasifikavimo matrica (žr. 28 lentelė), kurios bendras tikslumas – 0,34, o F_1 įvertis – 0,49. Nubraižius ROC kreivę, gautas 0,48 AUC įvertis (žr. 10 pav.). Beveik visi stebėjimai buvo priskirti vienai klasei.

28 lentelė. Klasifikavimo lentelė visiems duomenims su numatytaisiais parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	47	3	**Jautrumas**: 0,94
	0	93	3	**Specifiškumas**: 0,03
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0,5	**Bendras tikslumas**: 0,34



10 pav. ROC kreivė visiems duomenims su numatytaisiais parametrais.

Toliau buvo pritaikyta *GridsearchCV* funkcija geriausiam hiperparametrų rinkiniui rasti. Gautas rezultatas:

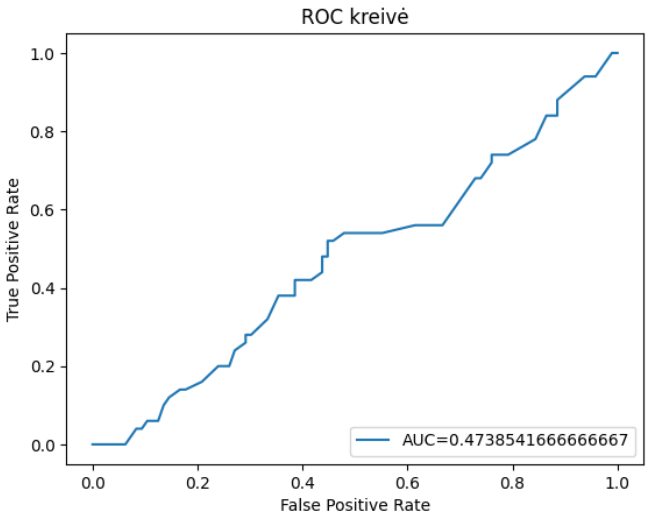
```
max_depth = 25, max_features = 0.5, n_estimators = 100, min_samples_split
= 15, random_state = 57
```

Su šiais hiperparametrais gautas klasifikavimo matricos bendras tikslumas – 0,53 (žr. 29 lentelė), o F_1 įvertis – 0,43. Gautas AUC įvertis – 0,47 (žr. 11 pav. ROC kreivė visiems duomenims su GridSearchCV parametrais.). Bendras tikslumas geresnis nei su numatytaisiais parametrais, tačiau AUC ir F_1 įverčiai prastesni.

29 lentelė. Klasifikavimo lentelė visiems duomenims su GridSearchCV parametrais.

		Prognozuoti	
		1	**0**

Tikri	**1**	26	24	**Jautrumas**: 0,52
	0	44	52	**Specifiškumas**: 0,54
		Tikslumas (precision): 0,37	**Neigiama prognostinė vertė**: 0,68	**Bendras tikslumas**: 0,53



11 pav. ROC kreivė visiems duomenims su GridSearchCV parametrais.

5 geriausi hiperparametrų rinkiniai gauti su *GridSearchCV*:

param_max_depth	param_max_features	param_min_samples_split	param_n_estimators
25	0.5	15	100
20	0.6	2	100
10	0.5	10	75
20	0.4	2	60
30	0.4	2	75

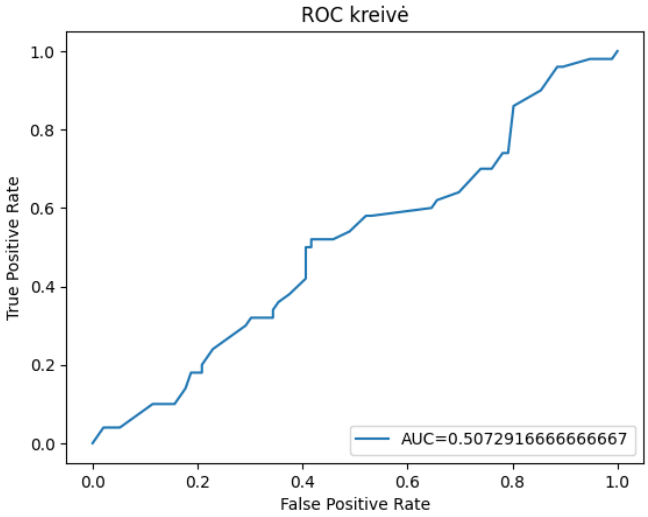
Tikslumas atitinkamai: 53,42; 47,94; 47,94; 40,41; 38,35. Galima pastebėti, kad tikslumas mažėja kartu su rinkinio reitingu. Taip pat bandymas rasti geriausius parametrus buvo atliktas ir pritaikius *RandomizedSearchCV* funkcija, kadangi tokiu būdu atsakymas gaunamas greičiau. Gautas hiperparametrų rinkinys:

```
random_state = 123, n_estimators = 75, min_samples_split = 10,
max_features = 0.6, max_depth = 20
```

Klasifikavimo matricos tikslumas bei F_1 įvertis gautas toks pat kaip su *GridSearchCV* funkcija, nors parametrai minimaliai skiriasi, tačiau AUC įvertis geresnis už abu ankstesnius gautus AUC įverčius su *GridSearchCV* ir numatytaisiais parametrais.

30 lentelė. Klasifikavimo lentelė visiems duomenims su RandomizedSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	26	24	**Jautrumas**: 0,52
	0	44	52	**Specifiškumas**: 0,54
		Tikslumas (precision): 0,37	**Neigiama prognostinė vertė**: 0,68	**Bendras tikslumas**: 0,53



12 pav. ROC kreivė visiems duomenims su RandomizedSearchCV parametrais.

Toliau pateikti 5 geriausi hiperparametrų rinkiniai gauti su RandomizedSearchCV.

param_random_state	param_n_estimators	param_min_samples_split	param_max_features	param_max_depth
123	75	10	0.6	20
50	75	10	0.5	30
50	75	5	0.4	15
50	75	2	0.4	30
123	75	5	0.4	15

Tikslumas atitinkamai: 53,42; 53,42; 34,24; 37,67; 43,83.

Tikslumas mažėja netolygiai – reitingas neatspindi prastesnio klasifikavimo tikslumo. Visam duomenų rinkiniui gautas geriausias bendras tikslumas (0,53) bei F_1 įvertis (0,43) su *GridSearchCV* ir *RandomizedSearchCV* gautais hiperparametrais. Geriausias AUC įvertis (0,5) gautas su *RandomizedSearchCV* parametrais.

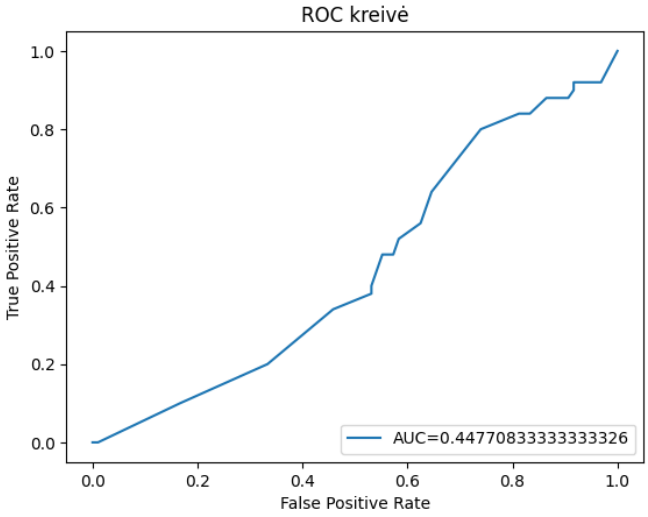
Reikšmingos kovariantės

Toliau klasifikavimas atliktas atrinktoms reikšmingoms kovariantėms – gliukozės kiekiui, neštumų skaičiui, KMI ir diabeto susirgimo funkcijai. Su numatytaisiais parametrais klasifikavimo matricos bendras tikslumas – 0,34 (žr. 31 lentelė), F_1 įvertis – 0,51, AUC – 0,45 (žr. 13 pav.). Galima pastebėti, kad visi stebėjimai yra priskirti vienai klasei.

31 lentelė. Klasifikavimo lentelė reikšmingoms kovariantėms su numatytais parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	50	0	**Jautrumas**: 1
	0	96	0	**Specifiškumas**: 0

	Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0	**Bendras tikslumas**: 0,34



13 pav. ROC kreivė reikšmingoms kovariantėms su numatytais parametrais.

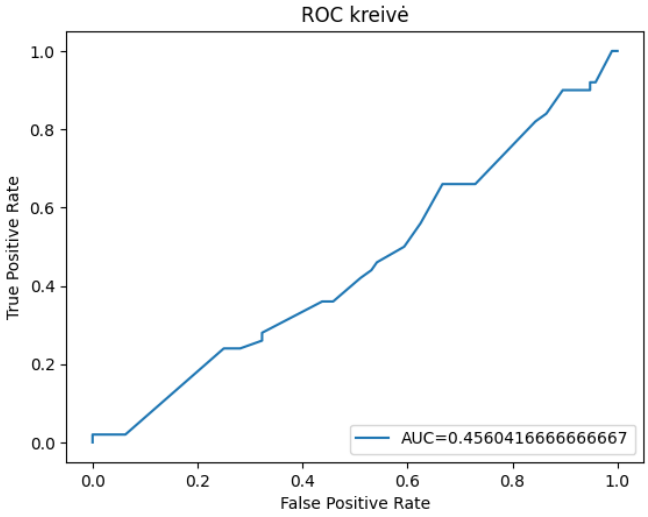
Su *GridSearchCV* atrinktas geriausias hiperparametrų rinkinys:

```
max_depth = 10, max_features = 0.4, n_estimators = 100, min_samples_split
= 2, random_state = 123
```

Visi stebėjimai vėl buvo priskirti vienai klasei ir gautas tikslumas bei F_1 įvertis toks pat kaip su numatytaisiais parametrais (žr. 32 lentelė), tačiau AUC įvertinimas aukštesnis (0,46) (žr. 14 pav.).

32 lentelė. Klasifikavimo lentelė reikšmingoms kovariantėms su GridSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	50	0	**Jautrumas**: 1
	0	96	0	**Specifiškumas**: 0
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0	**Bendras tikslumas**: 0,34



14 pav. ROC kreivė reikšmingoms kovariantėms su GridSearchCV parametrais.

5 griausi hiperparametrų rinkiniai gauti su *GridSearchCV*:

param_max_depth	param_max_features	param_min_samples_split	param_n_estimators	param_random_state
10	0.4	2	100	123
10	0.4	10	100	123
30	0.4	2	60	5
10	0.4	15	60	5
30	0.5	15	60	50

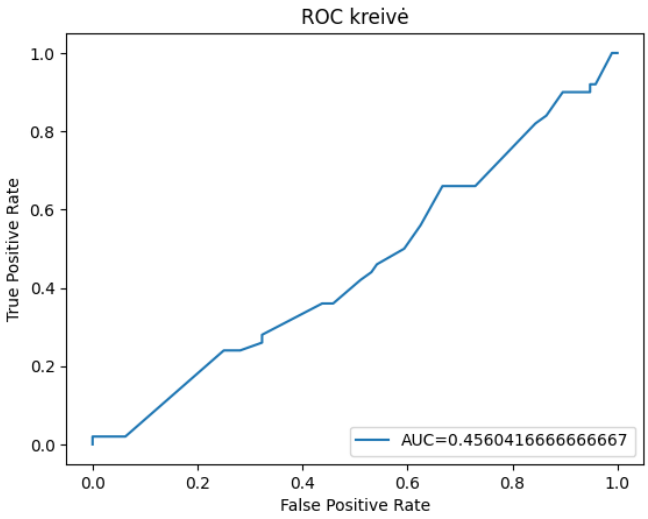
Visų tikslumas gautas vienodas – 34,25. Su *RandomizedSearchCV* gautas hiperparametrų rinkinys:

```
random_state = 123, n_estimators=100, min_samples_split = 5, max_features
= 0.5, max_depth = 25
```

Gautas rezultatas identiškai sutampa su *GridSearchCV* gautais rezultatais.

33 lentelė. Klasifikavimo lentelė reikšmingoms kovariantėms su RandomizedSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	50	0	**Jautrumas**: 1
	0	96	0	**Specifiškumas**: 0
		Tikslumas (precision): 0,34	**Neigiama prognostinė vertė**: 0	**Bendras tikslumas**: 0,34



15 pav. ROC kreivė reikšmingoms kovariantėms su RandomizedSearchCV parametrais.

5 geriausi hiperparametrų rinkiniai atrinkti su *RandomizedSearchCV*, kurių tikslumas nesiskyrė:

param_random_state	param_n_estimators	param_min_samples_split	param_max_features	param_max_depth
123	100	5	0.5	25
50	75	5	0.5	15
5	60	5	0.5	25
123	100	10	0.5	15
50	100	2	0.4	25

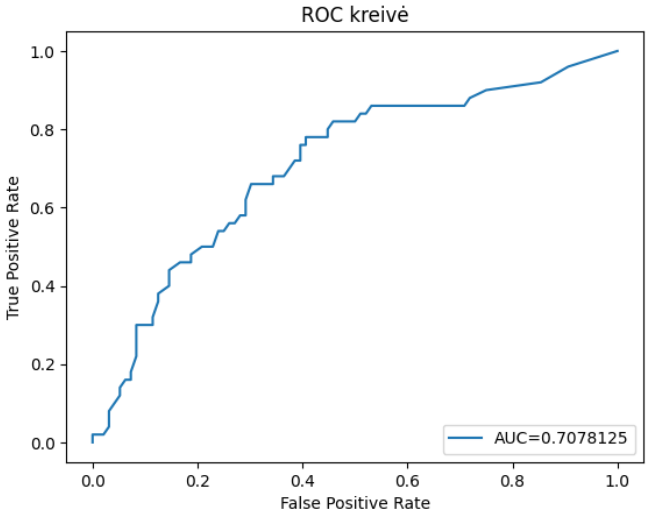
Taigi, reikšmingų kovariančių duomenų aibei rezultatai nebuvo žymiai geresni nei pilnam duomenų rinkiniui, visi stebėjimai buvo priskirti vienai klasei. Atrinkus hiperparametrus su *GridSearchCV* ir *RandomizedSearchCV* funkcijomis tikslumas nepagerėjo, bet padidėjo AUC įverčiai.

Sumažintos dimensijos

Duomenų aibė buvo sumažinta iki dviejų dimensijų pritaikius MDS euklidinį metodą ir jam buvo pritaikytas klasifikavimas. Su numatytaisiais parametrais gautas iki šiol geriausias bendras tikslumas – 0,69 (žr. 34 lentelė), F_1 įvertis smarkiai nepadidėjo – 0,53, tačiau ROC kreivė gauta labiausiai išsilenkusi ir AUC įvertis gana geras – 0,71 (žr. 16 pav.).

34 lentelė. Klasifikavimo lentelė sumažintos dimensijos duomenims su numatytais parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	25	25	**Jautrumas**: 0,5
	0	20	76	**Specifiškumas**: 0,79
		Tikslumas (precision): 0,56	**Neigiama prognostinė vertė**: 0,75	**Bendras tikslumas**: 0,69



16 pav. ROC kreivė sumažintos dimensijos duomenims su numatytais parametrais.

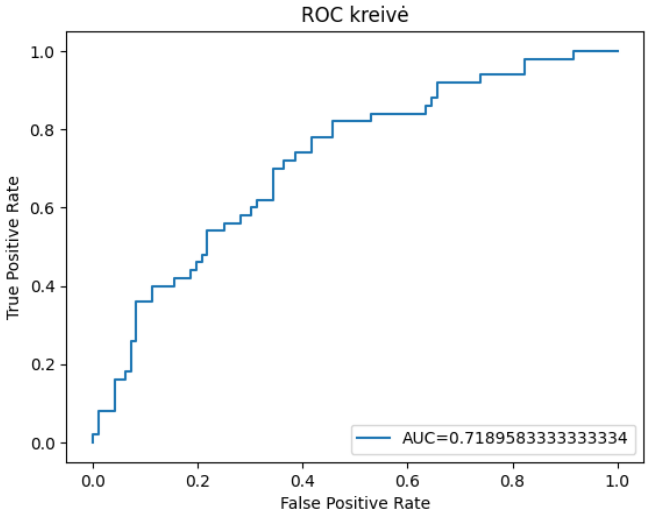
Tada buvo atrinktas geriausias hiperparametrų rinkinys su *GridSearchCV* funkcija:

```
max_depth = 10, max_features = 0.4, n_estimators = 60, min_samples_split =
2, random_state = 50
```

Pagal klasifikavimo lentelę galima matyti, kad bendras tikslumas (0,68) (žr. 35 lentelė) ir F_1 įvertis (0,51), nežymiai sumažėjo, bet gautas geresnis AUC įvertis – 0,72 (žr. 17 pav. ROC kreivė sumažintos dimensijos duomenims su GridSearchCV parametrais.).

35 lentelė. Klasifikavimo lentelė sumažintos dimensijos duomenims su GridSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	24	26	**Jautrumas**: 0,48
	0	20	76	**Specifiškumas**: 0,79
		Tikslumas (precision): 0,55	**Neigiama prognostinė vertė**: 0,75	**Bendras tikslumas**: 0,68



17 pav. ROC kreivė sumažintos dimensijos duomenims su GridSearchCV parametrais.

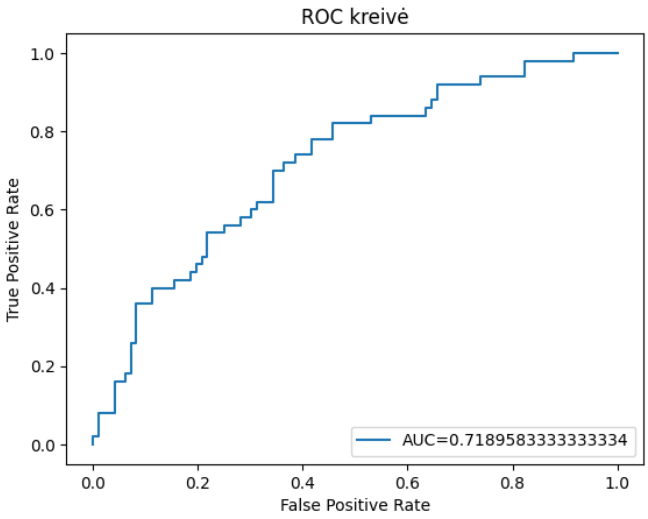
5 geriausi hiperparametrų rinkiniai gauti su *GridSearchCV*:

param_max_depth	param_max_features	param_min_samples_split	param_n_estimators	param_random_state
10	0.4	2	60	50
10	0.5	2	60	50
10	0.6	2	60	50
10	0.4	10	60	5
10	0.5	10	60	5

Pirmųjų trijų rinkinių bendras tikslumas bei klasifikavimo matricos sutapo, tačiau ketvirtas ir penktas rinkinys davė geresnį bendrą tikslumą – 0,71 (žr. 36 lentelė), tačiau AUC įvertis nepakito (žr. 18 pav. ROC kreivė sumažintos dimensijos duomenims su antrais GridSearchCV parametrais.18 pav.).

36 lentelė. Klasifikavimo lentelė sumažintos dimensijos duomenims su antrais GridSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	23	27	**Jautrumas**: 0,46
	0	15	81	**Specifiškumas**: 0,84
		Tikslumas (precision): 0,61	**Neigiama prognostinė vertė**: 0,75	**Bendras tikslumas**: 0,71



18 pav. ROC kreivė sumažintos dimensijos duomenims su antrais GridSearchCV parametrais.

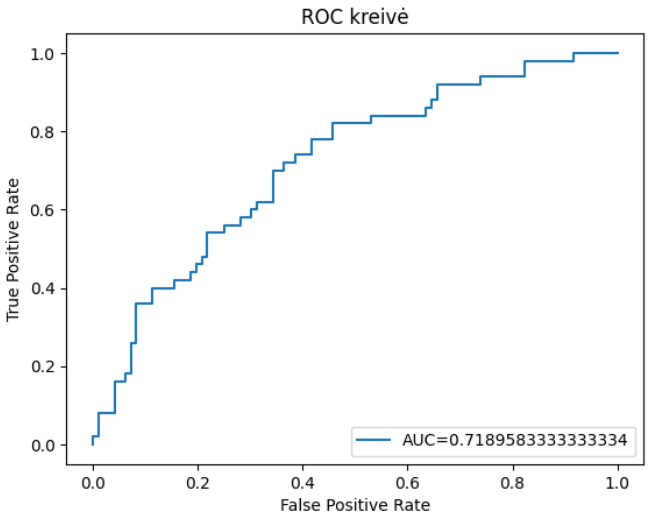
Su *RandomizedSearchCV* gautas geriausias hiperparamterų rinkinys:

```
random_state = 5, n_estimators = 100, min_samples_split = 2, max_features
= 0.6, max_depth = 10
```

Su *RandomizedSearchCV* gautas bendras tikslumas (0,68) (žr. 37 lentelė) bei AUC įvertis (0,72) (žr. 19 pav.) išliko toks pat, tačiau F_1 įvertis suprastėjo (0,49).

37 lentelė. Klasifikavimo lentelė sumažintos dimensijos duomenims su RandomiziedSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	22	28	**Jautrumas**: 0,44
	0	18	78	**Specifiškumas**: 0,81
		Tikslumas (precision): 0,55	**Neigiama prognostinė vertė**: 0,74	**Bendras tikslumas**: 0,68



19 pav. ROC kreivė sumažintos dimensijos duomenims su RandomizedSearchCV parametrais.

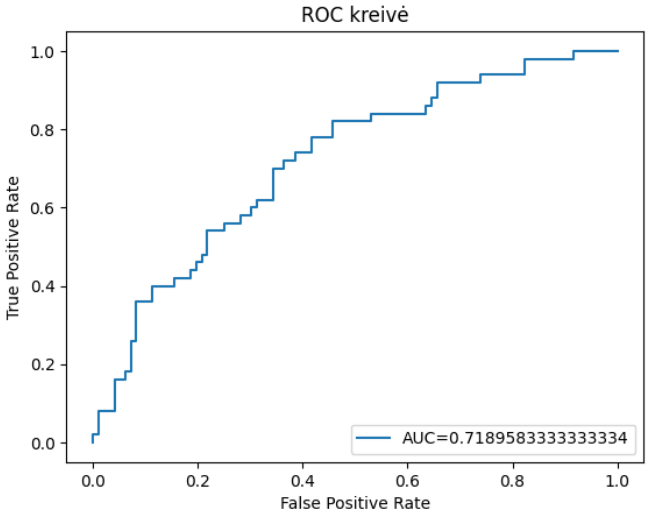
5 geriausi su *RandomSearchCV* atrinkti hiperparametrų rinkiniai:

param_random_state	param_n_estimators	param_min_samples_split	param_max_features	param_max_depth
5	100	2	0.6	10
5	100	10	0.6	25
5	100	5	0.6	15
5	75	10	0.6	30
50	75	5	0.4	10

Pirmojo rinkinio klasifikavimo rezultatai pateikti aukščiau. Antrojo rinkinio bendras tikslumas – 0,719, trečiojo – 0,698, ketvirtojo – 0,705, penktojo – 0,719. Toliau pateikta antrojo rinkinio klasifikavimo lentelė (žr. 38 lentelė) bei ROC kreivė. Nors bendras klasių klasifikavimo tikslumas yra geresnis nei pirmojo rinkinio, AUC įvertis išliko nepakitęs (žr. 20 pav.).

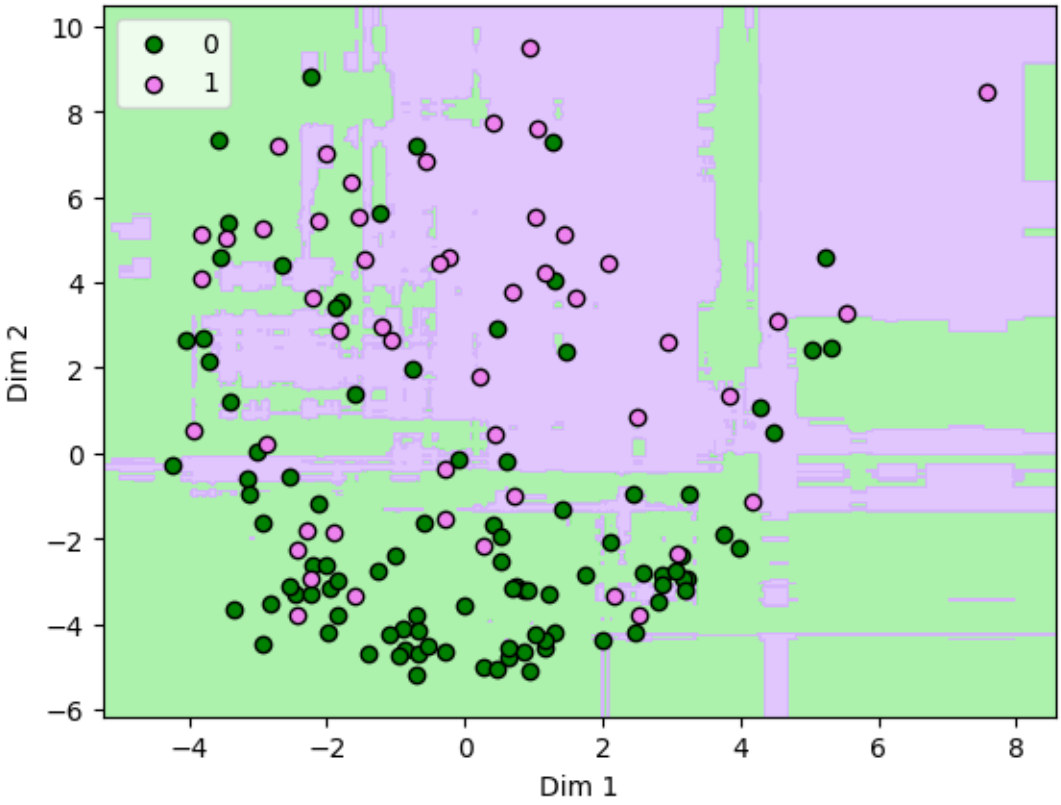
38 lentelė. Klasifikavimo lentelė sumažintos dimensijos duomenims su ketvirtais GridSearchCV parametrais.

		Prognozuoti		
		1	**0**	
Tikri	**1**	24	26	**Jautrumas**: 0,48
	0	15	81	**Specifiškumas**: 0,84
		Tikslumas (precision): 0,62	**Neigiama prognostinė vertė**: 0,76	**Bendras tikslumas**: 0,72



20 pav. ROC kreivė sumažintos dimensijos duomenims su ketvirtais GridSearchCV parametrais.

Toliau pateiktas sumažintos dimensijos su *GridSearchCV* atrinktais geriausiais hiperparametrais gautas klasifikavimas. Iš (21 pav.) galima matyti, kad dauguma taškų yra atitinkamos spalvos fone, reiškiančios, kad taškas buvo klasifikuotas teisingai.



21 pav. Sumažintos dimensijos su GridSearchCV parametrais duomenų klasifikavimas.

Taip pat iš pateiktos (39 lentelė) atsispindi visi atsitiktinio miško metodu gauti klasifikavimo rezultatai. Visiems požymiams aukštesniu AUC įverčiu išsiskyrė su *RandomizedSerachCV* gautų hiperparametrų rinkinio rezultatai. Reikšmingoms kovariantėms sutapo beveik visi rezultatai, tik su numatytaisiais AUC įvertis buvo žemesnis, bet nei vienas neparodė gero klasifikavimo. Tiksliausiai buvo klasifikuojamas sumažintos dimensijos duomenų rinkinys. Tikslumas smarkiai nesiskyrė visais atvejais F_1 įvertis prasčiausias buvo su *RandomizedSearchCV* atrinktais hiperparametrais, o su numatytaisiais parametrais gautas mažiausias AUC įvertis, todėl optimalus hiperparametrų rinkinys gautas su *GridSearchCV* funkcija.

39 lentelė. Atsitiktinio miško klasifikavimo rezultatai.

		Bendras tikslumas	F_1	**AUC**
Visi požymiai	Numatyti parametrai	0,34	0,49	0,48
	Grid SearchCV	0,53	0,43	0,47
	Randomized SearchCV	**0,53**	**0,43**	**0,5**
Reikšmingos kovariantės	Numatyti parametrai	0,34	0,51	0,45
	Grid SearchCV	**0,34**	**0,51**	**0,46**
	Randomized SearchCV	**0,34**	**0,51**	**0,46**
Sumažintos dimensijos	Numatyti parametrai	0,69	0,53	0,71
	Grid SearchCV	**0,68**	**0,51**	**0,72**
	Randomized SearchCV	0,68	0,49	0,72

GERIAUSIŲ KLASIFIKATORIŲ PALYGINIMAS

Geriausi visų 3 klasifikatorių rezultatai gauti naudojant sumažintos dimensijos duomenis.

40 lentelė. Geriausių klasifikatorių palyginimas sumažintos dimensijos duomenims

Naudotas klasifikatorius	**Parametrai**	**Bendras tikslumas**	F_1	**AUC**
Gauso Naivusis Bajeso	Neturi	0,75	0,59	0,70
Atsitiktinio miško	*{max_depth = 10, max_features = 0.4, n_estimators = 60, min_samples_split = 2, random_state = 50}*	0,68	0,51	0,72
Sprendimų medis	$\{Max\_depth: 4, max\_features: 0{,}4, min\_samples\_split: 5\}$	0,68	0,50	0,55

Iš (40 lentelė) matome, jog renkant geriausią klasifikatorių pagal bendrą tikslumą ir F_1 matą geriausiai pasirodė Gauso Naivusis Bajeso algoritmas (0,75 ir 0,59), o renkant pagal didžiausią plotą po ROC kreive – atsitiktinio miško klasifikatorius su parametrais: *{max_depth = 10, max_features = 0.4, n_estimators = 60, min_samples_split = 2, random_state = 50}*.

IŠVADOS

Duomenų aibę sudaro pacientų medicininiai požymiai. Klasifikavimui pasirinkta naudoti visus skaitinius požymius. Duomenys padalinti į mokymo ir testavimo aibes naudojant santykį 80 - 20, požymių matavimo skalės suvienodintos normuojant pagal min – max reikšmes. Klasifikavimui naudoti trys metodai: naivus Bajeso, sprendimų medis ir atsitiktinis miškas.

Reikšmingų kovariančių atrinkimas neturėjo teigiamos įtakos sprendimų medžio klasifikavimo rezultatams. Taip yra todėl, nes sprendimų medis savaime atlieka požymių atrinkimą konstruodamas sprendimų mazgus. Geriausi rezultatai gauti naudojant medžio gylį lygų 4 ir reikalaujant bent 5 stebėjimų norint toliau skaidyti duomenų aibę $\{Max\_depth: 4,\ min\_samples\_split: 5\}$ su sumažintos dimensijos aibe. Bendras modelio tikslumas gavosi 0,68.

Naudojant Naiviojo Bajeso klasifikavimo algoritmus geriausiai testiniai duomenys buvo suklasifikuoti sumažintos dimensijos aibėje, naudojant Gauso klasifikatorių: buvo pasiektas 0,75 bendras tikslumas, F_1 - 0,59, plotas po kreive 0,70, tikslumas (angl. Precision) buvo gautas 0,68. Suklasifikavus testinę duomenų aibę šiuo klasifikatoriumi: 12 stebėjimų buvo klaidingai priskirta diabeto liga, o 24 pacientams klaidingai liga nenustatyta testinėje aibėje. Šie taškai buvo ištirti išsamiau ir nustatyta, jog beveik visiems požymiams, išskyrus insulinui, klaidingų 1 medianos reikšmės buvo panašios, o kai kur net didesnės už tikrų 1. Taip pat amžiui, kraujo spaudimui, diabeto funkcijai, odos storiui medianų reikšmės buvo panašios klaidingų ir tikrų 0, o kitų požymių medianos reikšmės buvo didesnės nei tikrų 0. Visiems požymiams ir reikšmingiems Naiviojo Bajeso klasifikavimo algoritmai neparodė gerų rezultatų.

Pritaikius atsitikitinio miško metodą visiems požymiams aukštesniu AUC įverčiu (0,5) išsiskyrė su RandomizedSerachCV gautų hiperparametrų rinkinio rezultatai. Reikšmingoms kovariantėms sutapo beveik visi rezultatai, tik su numatytaisiais AUC įvertis buvo žemesnis (0,45), bet nei vienas neparodė gero klasifikavimo. Tiksliausiai buvo klasifikuojamas sumažintos dimensijos duomenų rinkinys. Tikslumas smarkiai nesiskyrė visais atvejais, F1 įvertis prasčiausias buvo su RandomizedSearchCV atrinktais hiperparametrais (0,49), o su numatytaisiais parametrais gautas mažiausias AUC įvertis (0,71), todėl optimalus hiperparametrų rinkinys gautas su GridSearchCV funkcija.

Geriausias klasifikatorius pagal bendrą tikslumą ir F_1 matą pasirodė Gauso Naivusis Bajeso algoritmas (0,75 ir 0,59), o pagal didžiausią plotą po

ROC kreive – atsitiktinio miško klasifikatorius su parametrais: *{max_depth = 10, max_features = 0.4, n_estimators = 60, min_samples_split = 2, random_state = 50}*.

LITERATŪRA IR ŠALTINIAI

[1] Sharma A., „Gaussian Naive Bayes with Hyperparameter Tuning“, 2021. https://www.analyticsvidhya.com/blog/2021/01/gaussian-naive-bayes-with-hyperpameter-tuning/

[2] „Turing“, „An Introduction to Naive Bayes Algorithm for Beginners“. https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners

[3] Yio T., „Understanding The Naive Bayes Classifier“, 2019. https://towardsdatascience.com/understanding-the-naive-bayes-classifier-16b6ee03ff7b#:~:text=Also%2C%20naive%20Bayes%20has%20almost,Bayes%20can%20be%20pretty%20inaccurate

[4] https://scikit-learn.org/stable/modules/naive_bayes.html

[5] https://datagy.io/sklearn-decision-tree-classifier/

[6] https://datagy.io/python-confusion-matrix/

[7] https://www.datacamp.com/tutorial/decision-tree-classification-python#

[8] https://www.ibm.com/topics/random-forest

[9] https://www.datacamp.com/tutorial/random-forests-classifier-python

[10] https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

[11] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html