Dimensijos mažinimas klasifikavime

Matas Gaulia, Vainius Gataveckas, Dovydas Martinkus
Duomenų Mokslas 3 kursas 2 gr.

Vilnius, 2022

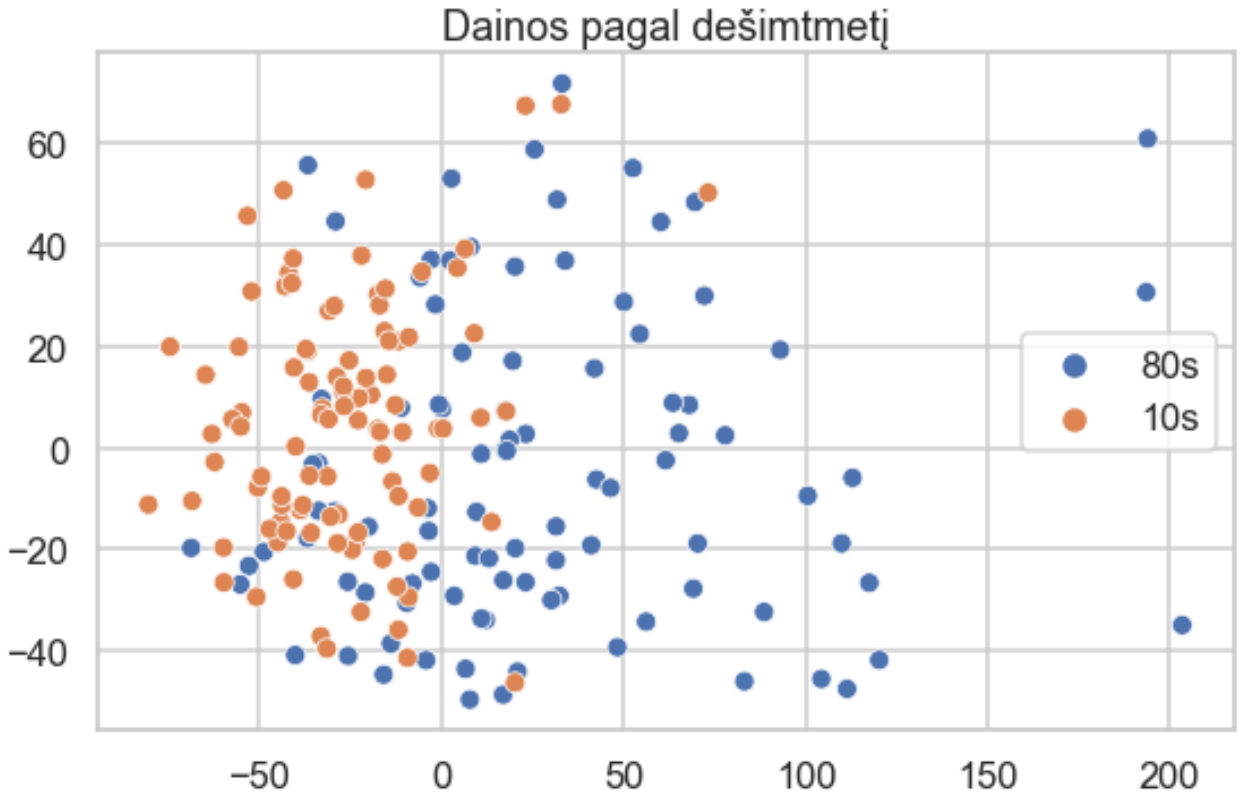
Naudoti duomenys

- decade - dainos sukūrimo metų dešimtmetis (80-ieji ar 2010-ieji)
- tempo - greitis
- energy - energiškumas
- danceability - šokamumas
- loudness – garsumas
- liveness - gyvumas
- valence – pozityvumas
- duration - trukmė
- acousticness - akustiškumas
- speechiness - žodžių kiekis dainoje
- popularity - populiarumas

Prieš tai laboratoriniuose naudotas Spotify dainų duomenų rinkinys.

Požymių matavimo skalės suvienodintos standartizuojant.

Šiai vizualizacijai dimensija sumažinta iki dim=2 naudojant PCA.



Naivus Bajeso klasifikatorius

- Optimalus požymių rinkiniai rasti naudojant kryžminę validaciją.
- Naivus Bajeso metodas neturi parametrų, kuriuos reiktų parinkti.
- Požymiai atrinkti godžiu algoritmu kiekviename žingsnyje šalinant tuo metu blogiausią požymį. Iš viso pašalintų požymių dalis *n_features_to_select* naudota kaip modelio parametras, kurio optimalios reikšmės ieškotas kryžminės validacijos būdu.

- Paryškinta – kryžmine validacija rasta optimali parametro reikšmė modeliui.
- *n_features_to_select*={0.2,0.4,**0.6**,0.8,1.0}.

- Toks gautas rezultatas yra natūralus, nes naivus Bajeso klasifikatorius priskiria vienodą svarbą visiems (ir mažiau informatyviems) požymiams.
- Optimaliam klasifikatoriui nenaudojami požymiai „Tempo“, „Energy“, „Liveness“, „Valance“.

Sprendimų medžio klasifikatorius

- Kadangi sprendimų medžiai požymį naudoja konstruoti sprendimų mazgui tik jeigu jis gerai atskiria klases (atlieka savaiminį požymių atrinkimą), todėl nesitikima gauti rezultatų pagerėjimo atrenkant požymių poaibį.

- *max_depth*={4,**5**,6},
- *min_samples_split*={2,**5**,10,15},
- *n_features_to_select*={0.6,0.8,**1.0**}

- Fiksavus kitų parametrų reikšmes, bet naudojant mažesnes *n_features_to_select* reikšmes dažniausiai gauti prastesni rezultatai lyginant su didesne požymių aibe. Tiesa, šie skirtumai maži.
- Parinkus optimalius parametrus stipriai pagerintas vidutinis kryžminės validacijos tikslumas lyginant su numatytaisiais parametrais.

Atsitiktinio miško klasifikatorius

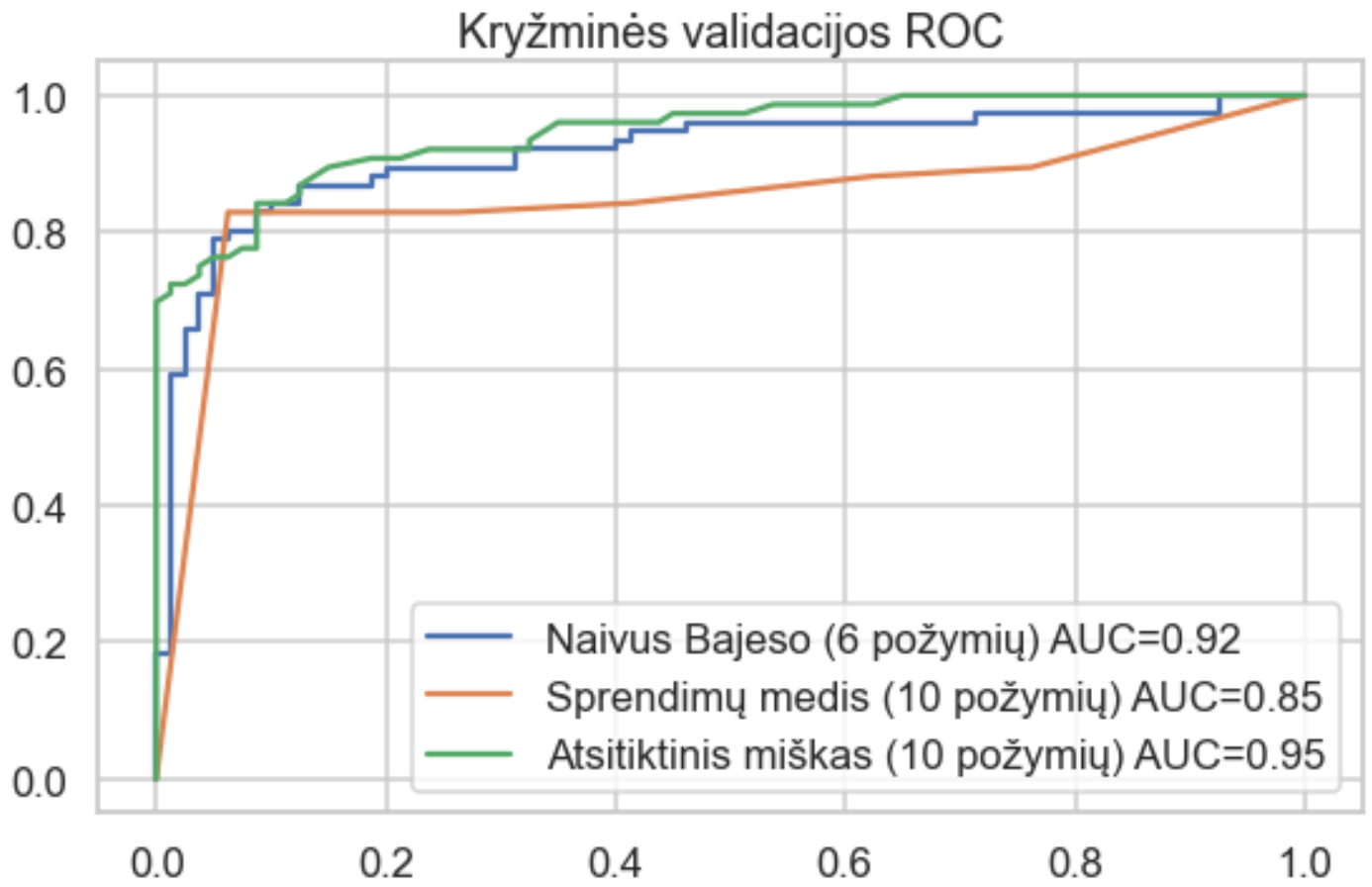
- Dėl atsitiktinumo atsitiktinio mokymo procese, kiekvieną kartą galima gauti kitą optimalių parametrų rinkinį, todėl prasmingą kalbėti tik apie geriausius parametrus fiksavus tam tikrą *random_state.*

- Kadangi metodas paremtas sprendimų medžiais, nesitikima gauti didėlės požymių šalinimo įtakos.

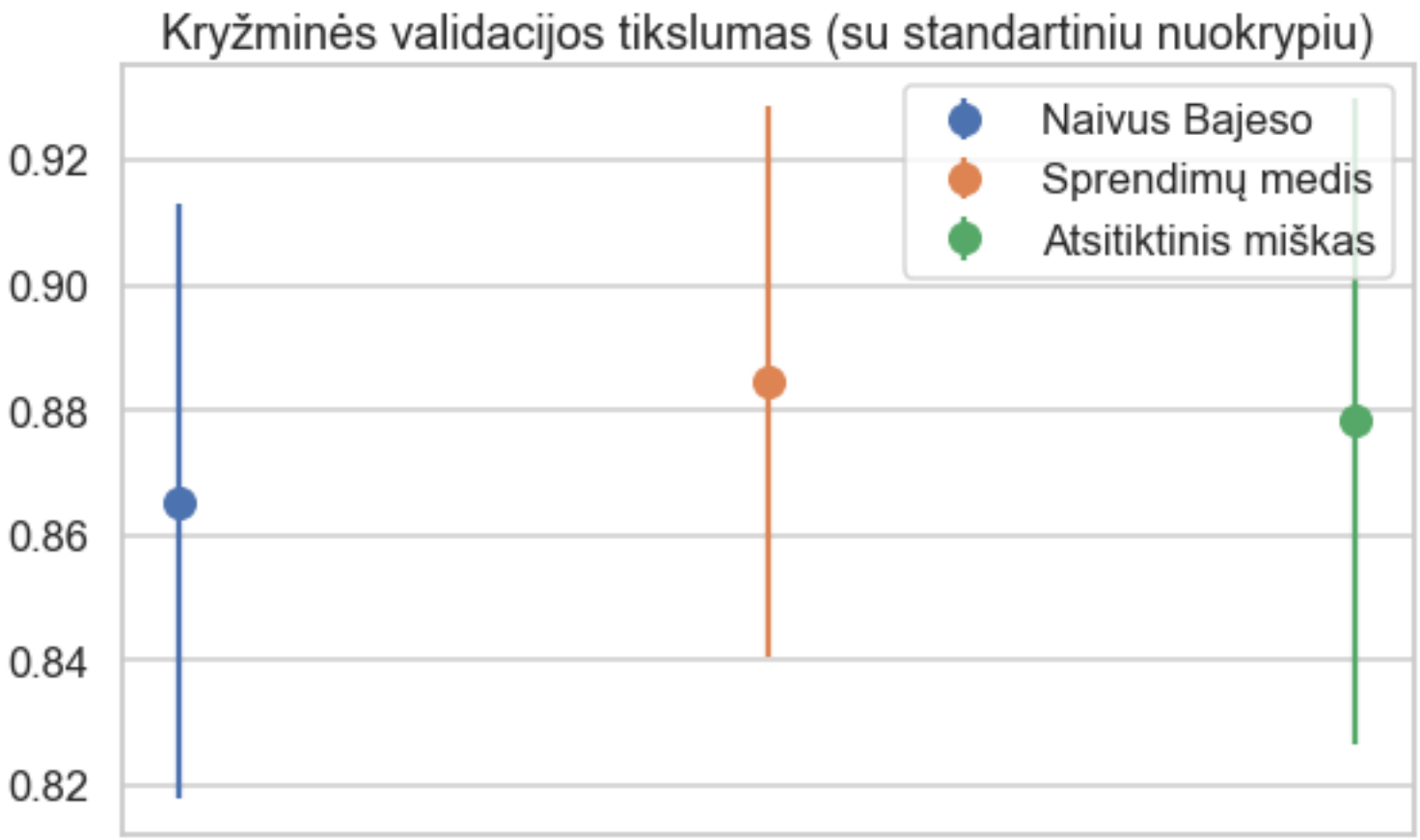
- n_estimators={25,50,**100**},
- *max_features*={**2**,3,4},
- *min_samples_split*={**2**,5,10,**15**},
- *n_features_to_select*={0.8,0.9,**1.0**}

- Tiek pats optimalus parametrų rinkinys, tiek su juo gautas vidutinis kryžminės validacijos tikslumas tik minimaliai skyrėsi nuo numatytųjų parametrų rezultatų.

Modelių palyginimas: ROC naudojant kryžminę validaciją



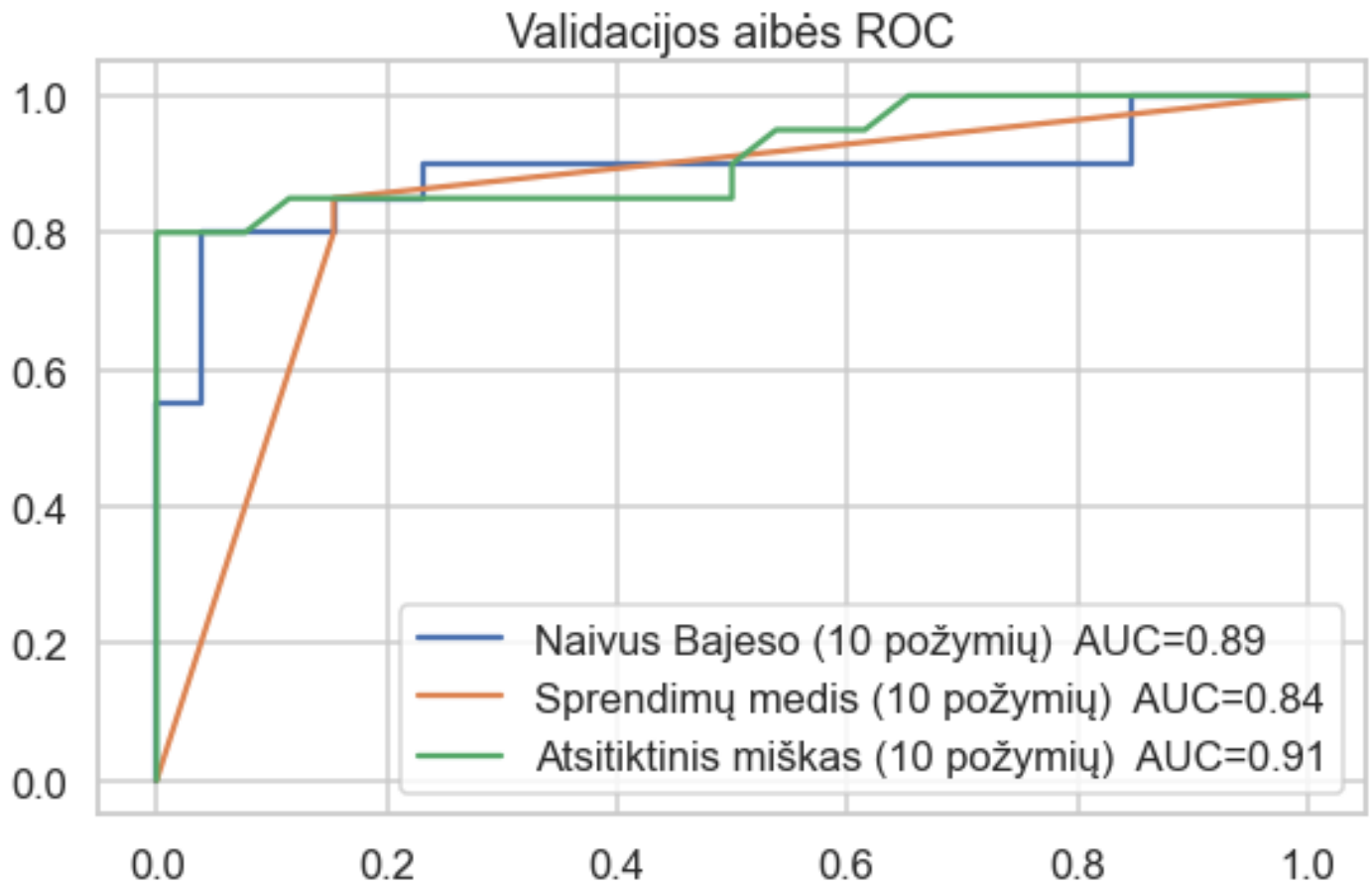
Modelių palyginimas: tikslumas naudojant kryžminę validaciją



Modelių palyginimas: kokybės matai naudojant kryžminę validaciją

Modelis	Klasė	Precision	Recall	F1-Score	Accuracy
Naivus Bajeso	10-ieji	0.85	0.90	0.87	0.87
Naivus Bajeso	80-ieji	0.89	0.83	0.86	0.87
Sprendimų medis	10-ieji	0.85	0.94	0.89	0.88
Sprendimų medis	80-ieji	0.93	0.83	0.88	0.88
Atsitiktinis miškas	10-ieji	0.86	0.91	0.89	0.88
Atsitiktinis miškas	80-ieji	0.90	0.84	0.87	0.88

Modelių palyginimas: ROC naudojant validacijos aibę

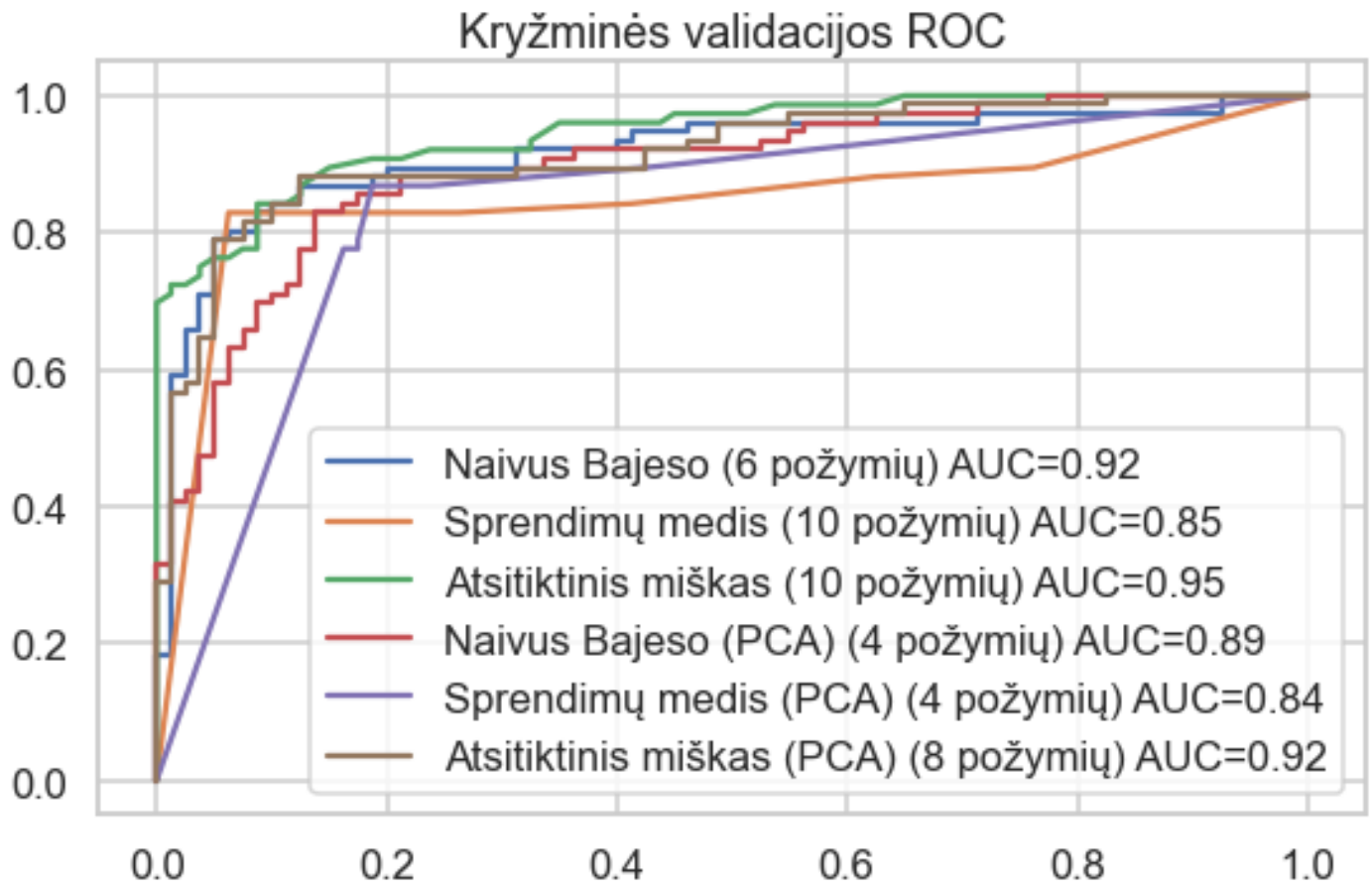


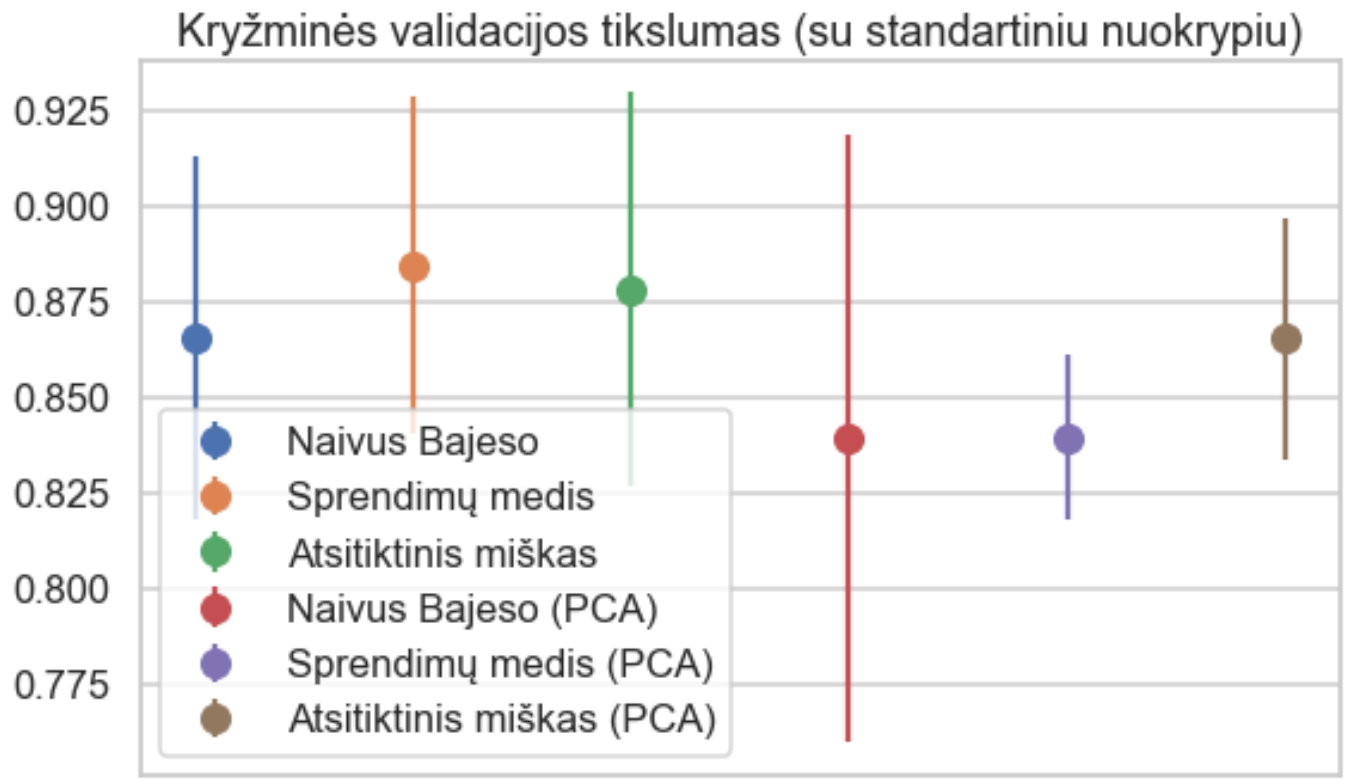
Modelių palyginimas: kokybės matai naudojant validacijos aibę

Modelis	Klasė	Precision	Recall	F1-Score	Accuracy
Naivus Bajeso	10-ieji	0.86	0.96	0.90	0.89
Naivus Bajeso	80-ieji	0.94	0.80	0.86	0.89
Sprendimų medis	10-ieji	0.88	0.85	0.83	0.85
Sprendimų medis	80-ieji	0.81	0.85	0.83	0.85
Atsitiktinis miškas	10-ieji	0.87	1.00	0.93	0.91
Atsitiktinis miškas	80-ieji	1.00	0.80	0.89	0.91

- Modelius lyginant naudojant kryžminę validaciją ir validacijos aibę matomi naivaus Bajeso ir atsitiktinio miško metodų pranašumas lyginant su sprendimų medžiu tiek pagal ROC grafiką, tiek pagal kitas modelio kokybės įvertinimo metrikas.

Modelių palyginimas: dimensijos mažinimo algoritmai



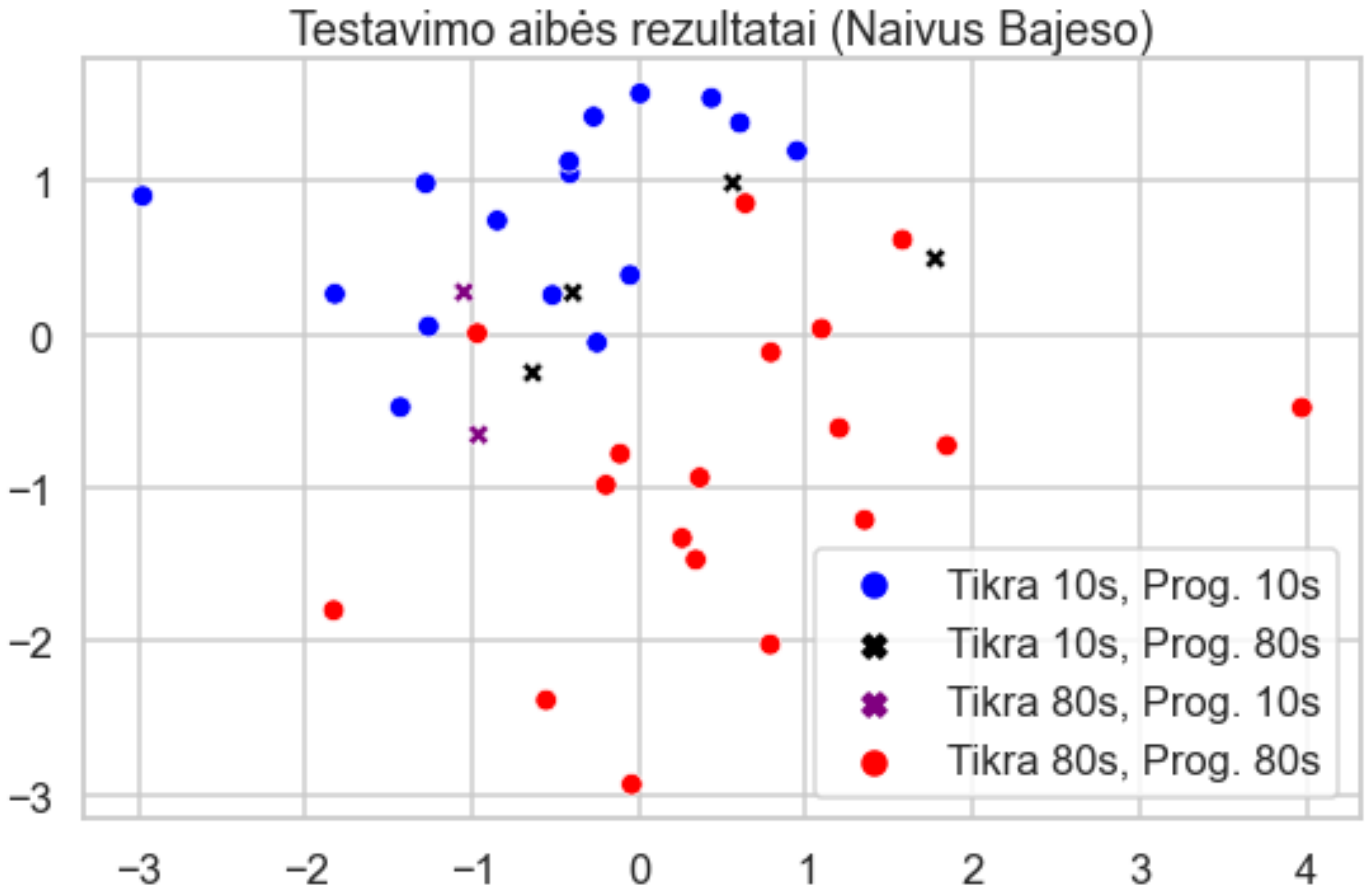

Kryžminės validacijos tikslumas (su standartiniu nuokrypiu)
0.925
0.900
0.875
0.850
0.825
0.800
0.775
Naivus Bajeso
Sprendimų medis
Atsitiktinis miškas
Naivus Bajeso (PCA)
Sprendimų medis (PCA)
Atsitiktinis miškas (PCA)

- Naudojant prieš tai aprašytą optimalių parametrų suradimo procedūrą, papildomai sudaryti modeliai, dimensijos mažinimui naudojantys PCA algoritmą vietoje rekursyvaus prasčiausių požymių eliminavimo.

- Visų trijų modelių atvejais pagal modelio kokybės metrikas matomi prastesni PCA metodo rezultatai.

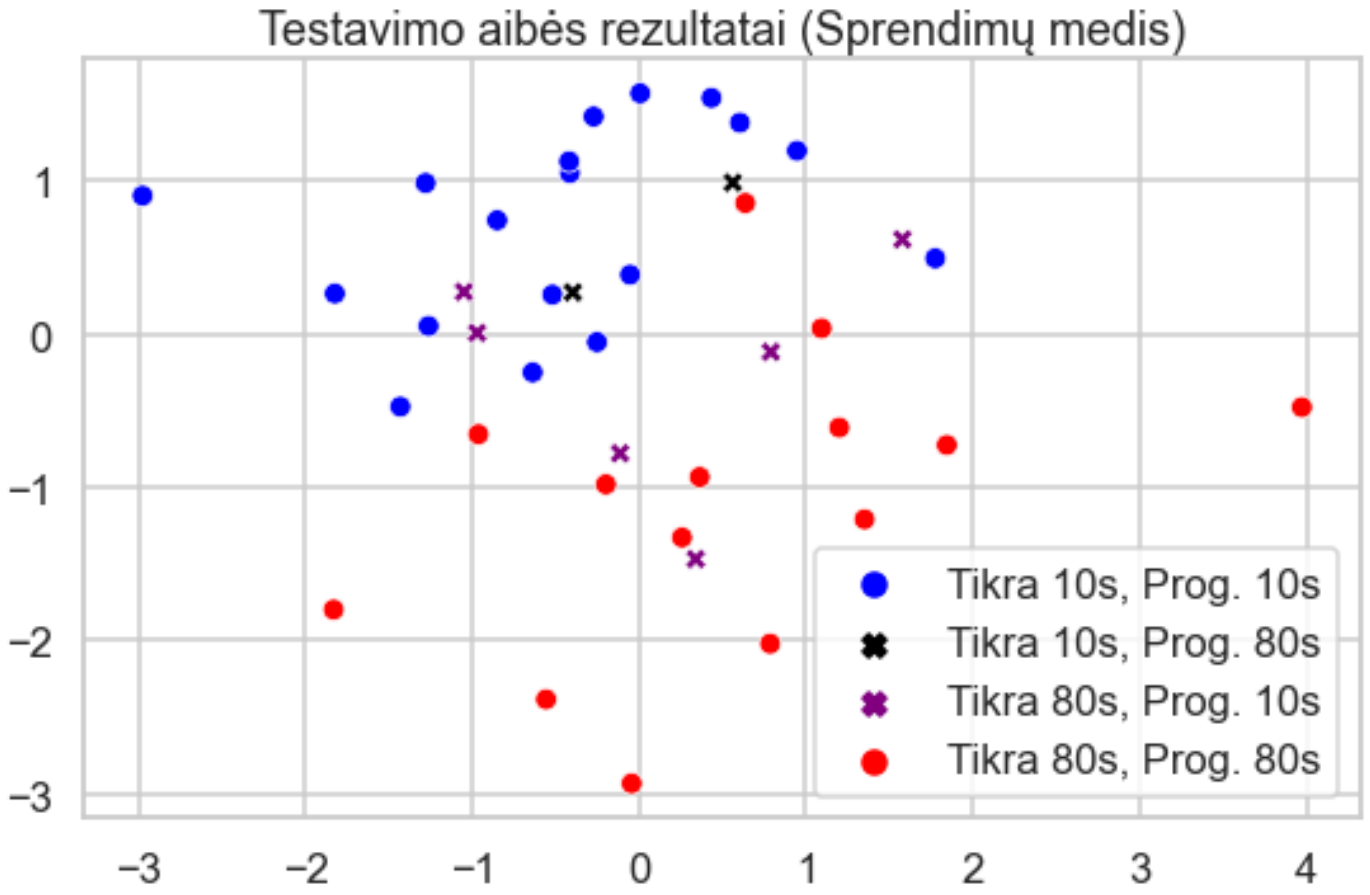
Naivus Bajeso rezultatai testavimo aibei

	Prognozuotos	
Tikros	16	4
	2	18



Sprendimų medžio rezultatai testavimo aibei

	Prognozuotos	
Tikros	18	2
	6	14



Atsitiktinio miško rezultatai testavimo aibei

	Prognozuotos	
Tikros	18	2
	1	19

- Pastebėta, kad blogai klasifikuotos dainos yra riboje tarp dviejų klasterių PCA metodu iki dim=2 sumažintoje erdvėje. Kai kurios dainos blogai klasifikuojamos visų trijų klasifikatorių.

Išvados

- Geriausi klasifikavimo rezultatai gauti naudojant atsitiktinio miško klasifikatorių. Metodas pasižymi ilgai trunkančia apmokymo trukme, tačiau šiuo atveju turima nesudėtinga duomenų aibė.

- Beveik tokie patys geri rezultatai gauti naudojant naivų Bajeso klasifikatorių. Šio klasifikatoriaus prielaidos yra visai natūralios turimoje duomenų aibėje. Modelis pasižymi aukštu mokymosi ir prognozavimo greičiu.

- Prasčiausi rezultatai gauti naudojant sprendimų medį.