Klasterizavimo algoritmai dideles apimties medicinos duomenims

Ataskaita už III-ąjį doktorantūros kursą
(2021/2022 metų II pusmetis)

Doktorantas: Roma Puronaitė

Vadovas: prof. dr. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2017-2023

2022 m. rugsėjis

Tyrimo objektas

- Didelės apimties medicinos duomenys, administracinio pobūdžio sveikatos duomenų rinkiniai
- Klasterizavimo algoritmai
- Prognostiniai depresijos prognozavimo modeliai

Tyrimo tikslas

- Pasiūlyti metodą didelės apimties medicinos duomenims klasterizuoti, atsižvelgiant į duomenų dinamikos laike savybes.
- Pasiūlyti prognostinį depresijos prognozavimo modelį, pritaikomą administracinio pobūdžio sveikatos duomenų rinkiniams.

Uždaviniai

- Ištirti dideles apimties medicinos duomenų klasterizavimui dažniausiai taikomus klasterizavimo metodus.
- Pasiūlyti klasterizavimo algoritmą ar esamo metodo patobulinimą, kuris atsižvelgtų į dinamiką laike.
- Pritaikyti atrinktus algoritmus ir pasiūlytą sprendimą realiems medicinos duomenų rinkiniams.
- Pasiūlyti algoritmo integravimo į sveikatos priežiūros įstaigos informacinę sistemą modelį.

Uždaviniai

- Ištirti dažniausiai taikomus prognostinius modelius depresijai prognozuoti.
- Pasiūlyti algoritmą ar esamo metodo patobulinimą, kuris būtų pritaikomas administracinio pobūdžio sveikatos duomenų rinkiniams ir atsižvelgtų į dinamiką laike.
- Pritaikyti pasiūlytą sprendimą realiems medicinos duomenų rinkiniams.
- Pasiūlyti algoritmo integravimo į sveikatos priežiūros įstaigos informacinę sistemą modelį.

Visos doktorantūros planas

Studijų metai	**Egzaminai**		**Dalyvavimas konferencijose**		**Publikacijos**		
	Planas	**Įvykdyta**	**Planas**	**Įvykdyta**	**Planas**	**Įvykdyta**	**Būklė** [4]
I (2017/2018)	2	2		1		1	1 publikuota
II (2018/2019)	2	0	1	5		2	2 publikuota
III (2021/2022)	**0**	**2 (skola iš II metų)**	**1**	**2**	**1**	**2**	**2 publikuota**
IV (2022/2023)	0	0			*1*		
Iš viso:	4	4	2	8	2	5	

Ataskaita už III mokslo metus, 2 pusmetį

- Egzaminai

Egzaminai		
Planas	**Įvykdyta**	**Būklė**
Netiesiniai statistikos modeliai masinių duomenų analizėje	Netiesiniai statistikos modeliai masinių duomenų analizėje, 2022 m. kovo 18 d.	Išlaikytas
Daugiamačių duomenų vizualizavimo metodai	Daugiamačių duomenų vizualizavimo metodai, 2022 m. kovo 30 d.	Išlaikytas

- Konferencijos

Dalyvavimas konferencijose		
31st International Biometric Conference, 2022 m. liepos 10-15 d., Ryga, Latvija	**Puronaite, Roma**, Ramanauskaitė, Dovilė, Burneikaitė, Greta, Švaikevičienė, Kristina, Šavareikaitė, Alicija, Vaitkute, Samanta, Jakaitienė, Audronė, Dambrauskas, Laimis, Jurevičienė, Elena, Trinkūnas, Justas, Kasiulevičius, Vytautas, Kazėnaitė, Edita, „Challenges of modeling depression and anxiety risk using data from large healthcare databases: systematic review and situation analysis“, *31st International Biometric Conference*, 2022 m. liepos 10-15 d., Ryga, Latvija	Tarptautinė

- Publikacijos

Publikacijos			
Planas	**Įvykdyta**	**Būklė**	**Publikacijos tipas**
	Paskutinį pusmetį naujų publikacijų nėra.		

Mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
1	Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje): Anotuotos bibliografijos sudarymas. Mokslinės literatūros apžvalga. Egzistuojančių metodų taikymo medicinos duomenims analizavimas.	2017 m. spalio mėn. – 2018 m. birželio mėn.	Naudojant sisteminės literatūros apžvalgos metodą, atlikta literatūros analizė šiomis temomis: Sisteminė literatūros apžvalga – dauginių ligų modelių klasterizavimo metodai (išnagrinėta 151 publikacija, identifikuoti šioje tematikoje naudojami klasterizavimo metodai) ir Sisteminė literatūros apžvalga – netiesiniai statistikos modeliai: dauginių ligų analizės tyrimai naudojant didelės apimties medicinos duomenų masyvus (išnagrinėta 206 publikacijos, identifikuoti šioje tematikoje naudojami klasterizavimo metodai) Apžvelgtos antriniu tikslu naudojamos administracinės duomenų bazės analizės strategijos
2	Mokslinio tyrimo vykdymas: 2.1. Tyrimo metodikos sudarymas: 1. Disertacijos tikslo formulavimas. 2. Disertacijos uždavinių formulavimas.	2018 m. birželio mėn. – 2018 m. lapkričio mėn.	Atlikus literatūros analizę nustatyta, kad tokio tipo duomenys (elektroninių sveikatos įrašų, administracinių duomenų bazių) dažniausiai analizuojami taikant skerspjūvio tipo metodus taikant klasterizavimo metodus duomenims tam tikrame stebėjimo taške ar apibendrintai tam tikro laikotarpio informacijai, neatsižvelgiant į dinamiką. Nuspręsta nagrinėti nuo laiko priklausomų sąryšių įvertinimo galimybes. Disertacijos tikslas: Pasiūlyti klasterizavimo algoritmų papildymą didelės apimties medicinos duomenims, atsižvelgiant į duomenų dinamikos laike savybes. Disertacijos uždaviniai: 1. Ištirti didelės apimties medicinos duomenų klasterizavimui dažniausiai taikomus klasterizavimo metodus. 2. Pasiūlyti klasterizavimo algoritmą ar esamo metodo patobulinimą, kuris atsižvelgtų į dinamiką laike. 3. Pritaikyti atrinktus algoritmus ir pasiūlytą sprendimą realiems medicinos duomenų rinkiniams. 4. Pasiūlyti algoritmo integravimo į sveikatos priežiūros įstaigos informacinę sistemą modelį.

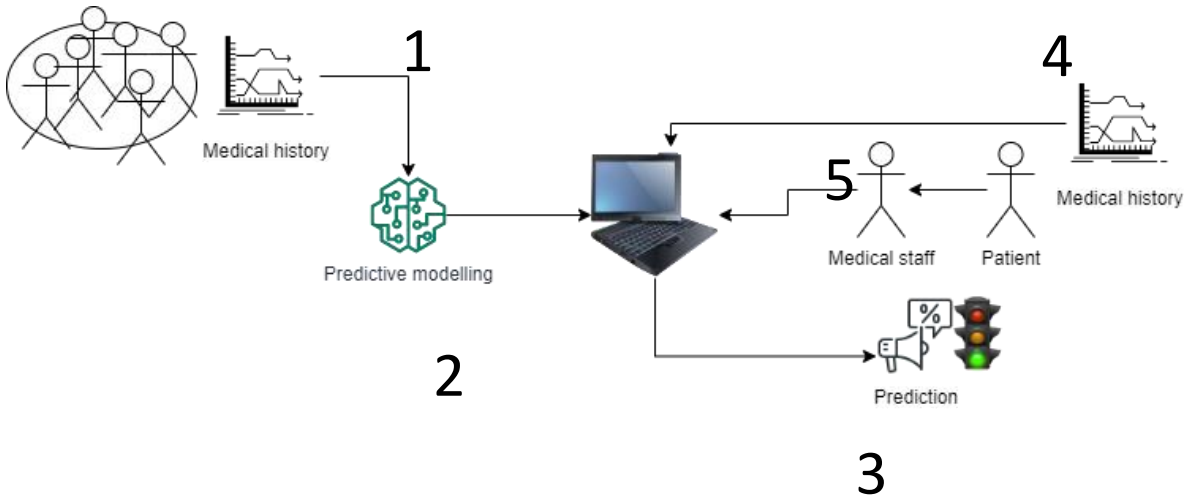
	2.2. Teorinis tyrimas: Matematinio modelio sudarymas. Algoritmų konstravimas ir tobulinimas.	2018 m. lapkričio mėn. – 2019 m. birželio mėn.	Nagrinėjami ir išbandomi nagrinėtoje literatūroje pasiūlyti metodai atsižvelgiantys į duomenų dinamiką. Empiriniui tyrimui atrinkta: • Skerspjūvio tipo analizė • Temporalinio fenotipavimo metodai (PARAFAC2, neneigiamas matricos faktorizavimas) • Laiko dimensijos transformavimas (siūloma transformacijai naudoti Poincare grafikus (indeksus), laiko eilučių išlyginimą, netiesinius statistinius modelius) • Kiti: gilusis mokymas, gilieji rekurentiniai neuroniniai tinklai, arba giliojo mokymo teksto analizės metodai.
	2.3. Empirinis tyrimas: 1. Siūlomų algoritmų pritaikymas medicinos duomenims. 2. Siūlomų algoritmų tobulinimas, atsižvelgiant į gautus rezultatus.	2018 m. birželio mėn. – 2020 m. vasario mėn. **2018 m. birželio mėn. – 2022 m. vasario mėn.***	Dauginių ligų duomenys pagal atitinkamus ligų kodus buvo suskaidyti į pacientų turinčių tam tikrą ligą pogrupius: • Diabetas • Lėtinė obstrukcinė plaučių liga • Depresija ir nerimas Kiekvieno paciento atveju buvo sudaryti dvinariai kintamieji atitinkantys ligos (išskirta 30 lėtinių ligų) buvimą arba nebuvimą atitinkamo paciento ligos istorijoje per tyrimo laikotarpį. Šiuose pogrupiuose pritaikius hierarchinį klasterizavimą (angl. HCA) gauti ligų klasteriai, kurie buvo vertinti gydytojų specialistų, dalis šių klasterių įvertinti kaip turintys klinikinį paaiškinimą. Taip pat pritaikyta tiriamoji faktorinė analizė (angl. EFA). Koreliacijos matrica sudaryta naudojant tetrachorinę koreliaciją tarp dvinarių kintamųjų porų. Atlikta vizuali daugiamačių dauginių ligų duomenų analizė Pradėta: • Temporalinio fenotipavimo metodai (PARAFAC2, neneigiamas matricos faktorizavimas) Numatoma: • Laiko dimensijos transformavimas (siūloma transformacijai naudoti Poincare grafikus (indeksus), laiko eilučių išlyginimą, netiesinius statistinius modelius) • Kiti: gilusis mokymas, gilieji rekurentiniai neuroniniai tinklai, arba giliojo mokymo teksto analizės metodai.

	2.4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas: 1. Algoritmų tikslumo įvertinimas, palyginimas su kitų autorių metodais, atrinktais remiantis išanalizuota moksline literatūra. 2. Gautų rezultatų apibendrinimas. 3. Išvadų parengimas.	2020 m. vasario mėn. – 2020 m. liepos mėn. **2022 m. vasario mėn. – 2022 m. liepos mėn.***	2021/2022 metais I pusmetis Planuojamos publikacijos (pradėta rengti): Depresijos ir nerimo prognozavimas remiantis administracinio pobūdžio duomenimis, pritaikant Poincare plot metodą ir apskaičiuotus indeksus (Health Informatics tematikos žurnale) 2021/2022 metais II pusmetis Pranešimas konferencijoje: **Puronaite, Roma**, Ramanauskaitė, Dovilė, Burneikaitė, Greta, Švaikevičienė, Kristina, Šavareikaitė, Alicija, Vaitkute, Samanta, Jakaitienė, Audronė, Dambrauskas, Laimis, Jurevičienė, Elena, Trinkūnas, Justas, Kasiulevičius, Vytautas, Kazėnaitė, Edita, „Challenges of modeling depression and anxiety risk using data from large healthcare databases: systematic review and situation analysis“, *31st International Biometric Conference*, 2022 m. liepos 10-15 d., Ryga, Latvija Vykdomi: Prognostinių modelių depresijai prognozuoti eksperimentai.
3	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas: 1. Tyrimų apžvalga ir analizė. 2. Tyrimo metodikos sudarymas. 3. Teorinis tyrimas. 4. Empirinis tyrimas. 5. Gautų duomenų analizė, apibendrinimas. 6. Išvados, įvadas, literatūros sąrašas.	2020 m. rugsėjo mėn. – 2021 m. gegužės mėn. **2022 m. rugsėjo mėn. – 2023 m. gegužės mėn.***	2021/2022 metais I pusmetis Pradėtos rengti dalys: mokslinės literatūros apžvalga, tyrimo metodika, duomenų analizė. 2021/2022 metais II pusmetis Tęsiama: mokslinės literatūros apžvalga, tyrimo metodika, duomenų analizė. Pradėtos rengti dalys: eksperimentinė dalis.

4	Daktaro disertacijos parengimas ir svarstymas padalinyje	2021 m. birželio mėn. **2023 m. birželio mėn.***	
5	Daktaro disertacijos gynimas	2021 m. rugsėjo mėn. **2023 m. rugsėjo mėn.***	

IBC2022 konferencija
Sisteminė literatūros apžvalga

Tikslas: taikant sisteminės literatūros apžvalgos metodą
išrinkti ir apibendrinti informaciją apie depresijai ir nerimui
prognozuoti taikomus metodus.

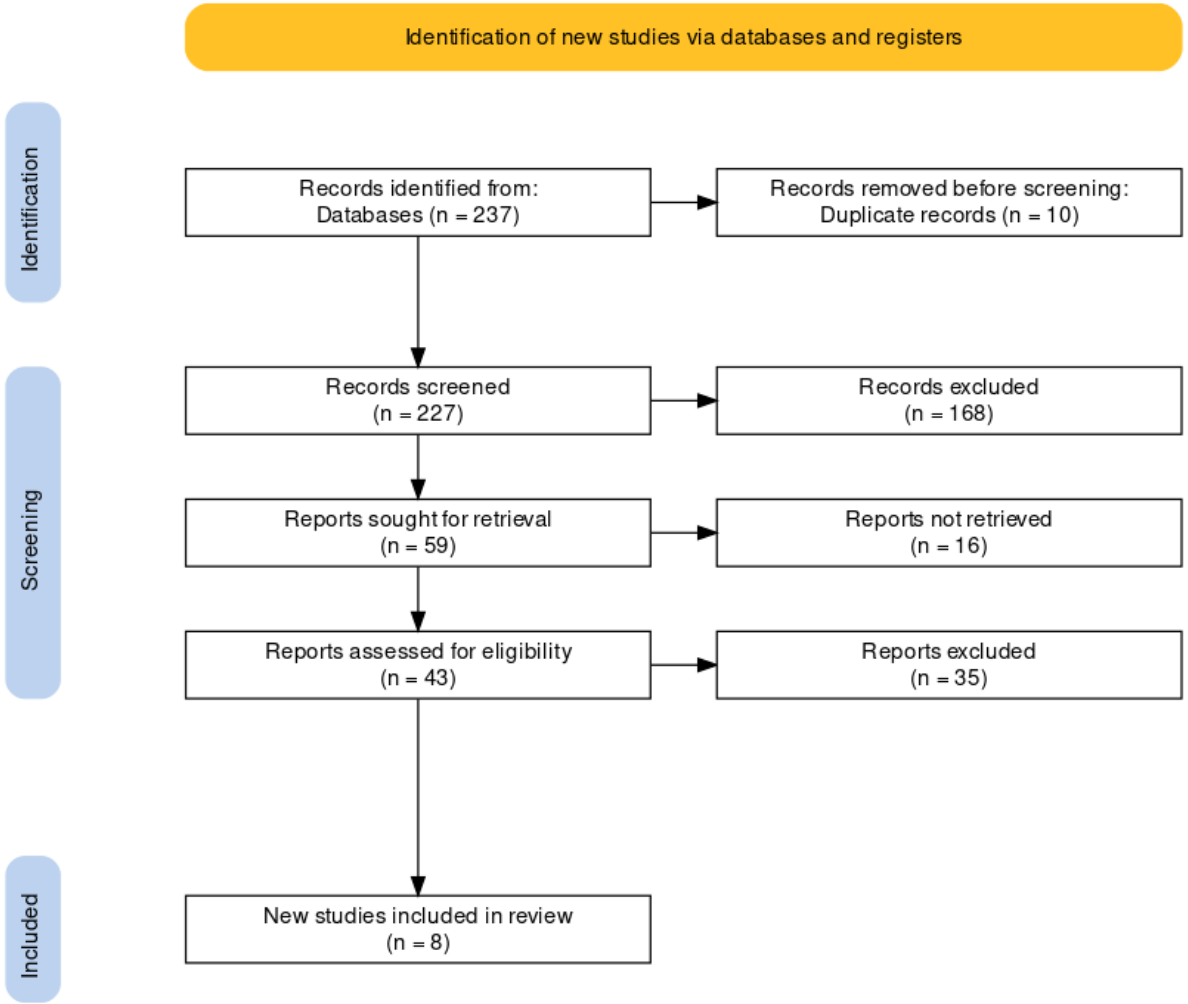


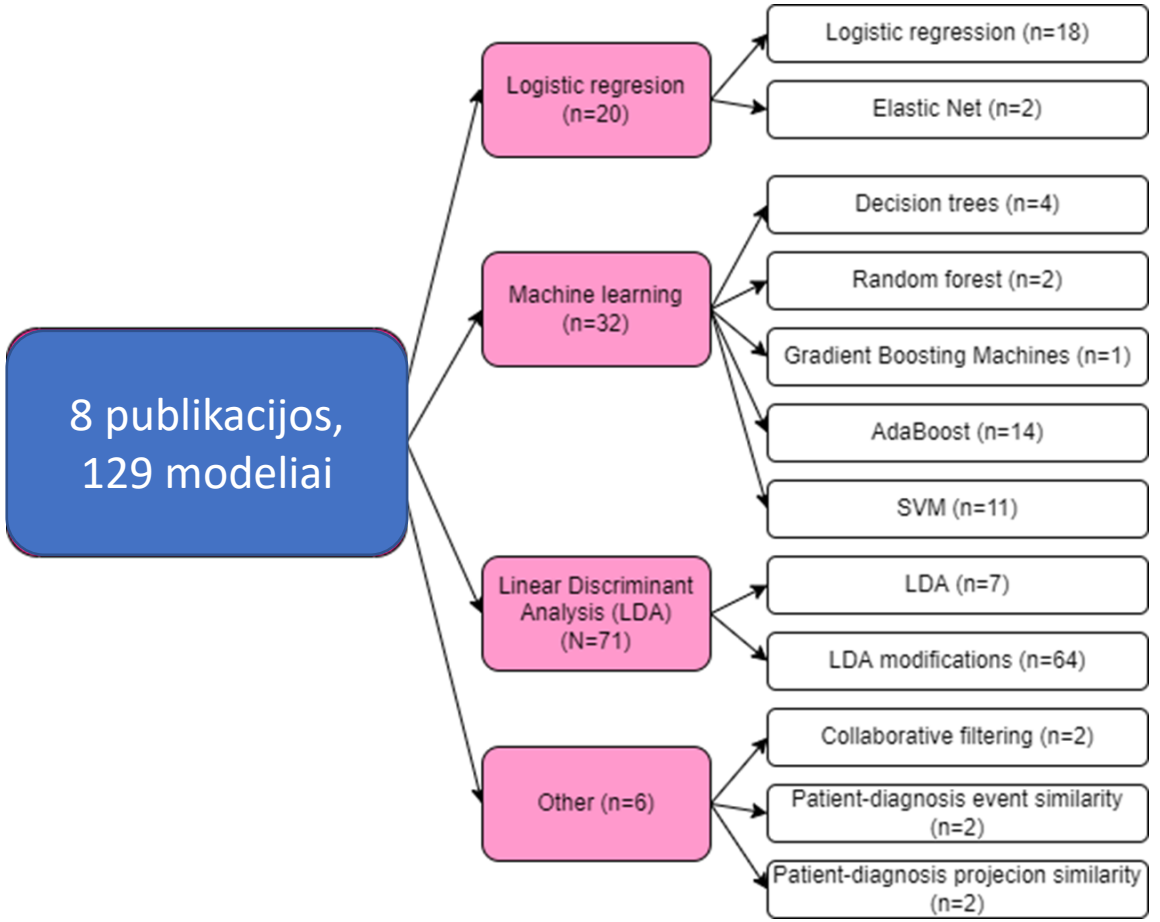
Pagrindiniai tyrimo tikslai apžvelgti šiuos prognostinio modelio kūrimo etapus:

1. Duomenų tvarkymas ir parengimas
2. Modelių parinkimas
3. Modelių tikslumo vertimas
4. Vidinis ir išorinis validumas
5. Praktinis pritaikymas

Priemonės: CHARMS klausimynas (publikacijų sisteminimui), TRIPOD klausimynas (atitiktis gerajai prognostinių modelių aprašymo praktikai), PROBAST klausimynas (šališkumo rizikos įvertinimas).

IBC2022 konferencija
Sisteminė literatūros apžvalga





Apibendrinimas

- Modeliai depresijai prognozuoti:

- Vidutinis AUC 0,646
- Geriausias modelis: Logistinė regresija, AUC = 0,782

- 5 publikacijose nurodoma, kad vertintas vidinis validumas, nei viena nenurodo išorinio validumo vertinimo.
- Hiperparametrų parinkimui naudojamas CV metodas (nenaudojama atskira validavimo imtis.

- Atitikimas TRIPOD neviršija 75 proc. (max. 100 proc., vertinami 29 teiginiai).
- Mažiausias atitikimas nustatytas aprašant prediktorius, rezultatus, pritaikomumą praktikoje.
- PROBAST: 5 iš 8 – aukštos šališkumo rizikos, 1 – žemos, 2 – neužtenka duomenų įvertinti.
- Daugiausiai aukštai rizikai priskiriama dalis: pacientų grupės atrankos ir aprašymo.

Eksperimentai: Prognostinio modelio kūrimas

- Analizuojama populiacija: Lėtinėmis ligomis, įvertintomis vienerių metų ligos istorija, sergantys asmenys, kuriems vienerius metus nebuvo nustatyta depresijos diagnozė.
- Grupė 0 – asmenys iš analizuojamos populiacijos, kurie per ateinančius metus nuo vertinimo taško (index) neturi depresijos diagnozės.
- Grupė 1– asmenys iš analizuojamos populiacijos, kurie per ateinančius metus nuo vertinimo taško (index) turi depresijos diagnozę (F32.* arba F33.*).

Eksperimentai: Prognostinio modelio kūrimas

1. Duomenų rinkiniai: DEP12_plus_DGN ir DEP12_plus_DGN_ATC.

DGN: 31 lėtinė liga pagal lėtinių ligų apibrėžimą (remiantis tarptautinės ligų klasifikacijos kodais, pvz. Diabetas: E10, E11).

ATC: 108 vaistų kodai nurodantys vaistų terapinę grupę (pvz. gliukozę mažinantys vaistai: A10B

2. Mokymo ir testavimo rinkinius (70/30).

Mokymo imtis: DEP0: 238092; DEP1: 3124; 1,3 proc.

Testavimo imtis: DEP0: 101997; DEP1: 1380; 1,3 proc.

Eksperimentai: Prognostinio modelio kūrimas

3. Mokymo duomenų rinkinio klasių subalansavimas: dirbtinė mažumos sukūrimo technika (Synthetic Minority Oversampling Technique, SMOTE), prisitaikančių kaimynų dirbtinė mažumos sukūrimo technika (angl. Adaptive Neighbor Synthetic Minority Oversampling Technique, ANS).

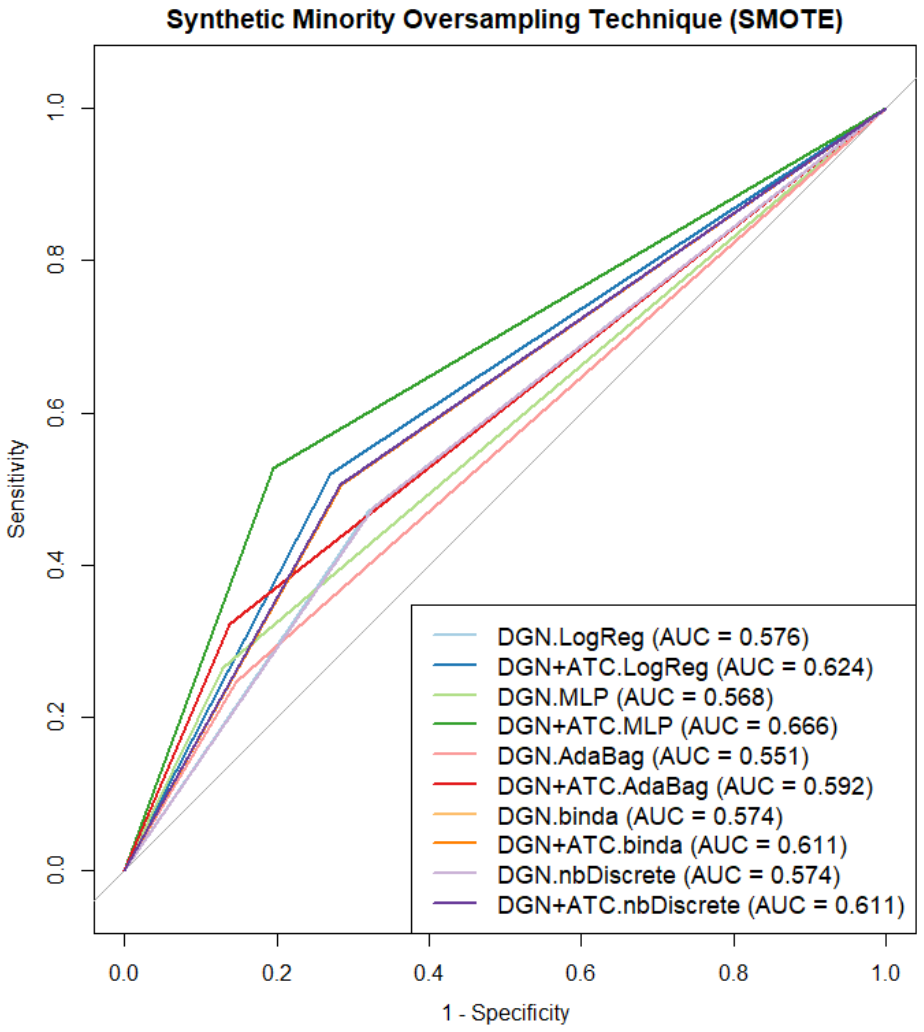
4. Prediktorių atrinkimas (feature selection). Chi-square of independence test, $p < 0,05$

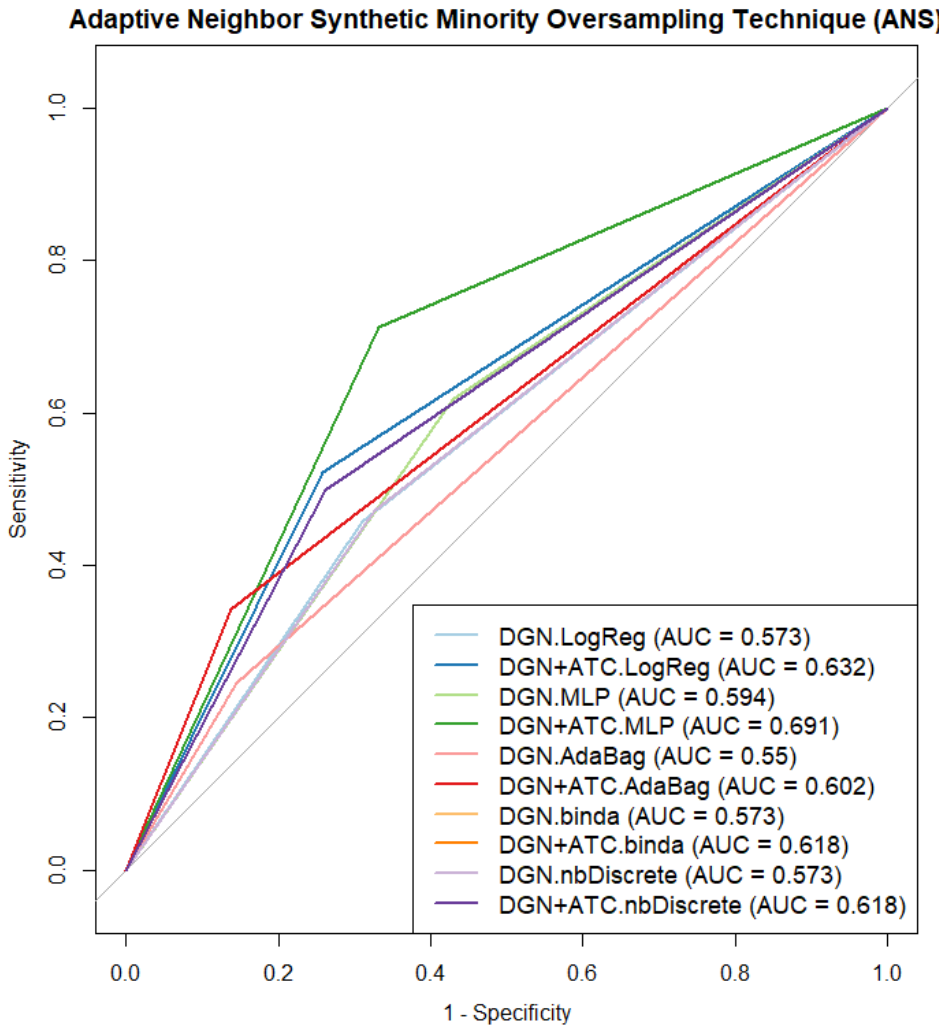
5. Atliekamas modeliavimas pritaikant kelis skirtingus metodus:

logistinės regresijos (LogReg), daugiasluoksnis perceptronas (MLP), AdaBag, diskriminantinės analizės (binda), Naivaus Bajeso (nbDiscrete).

6. Hiperparametrų parinkimui naudojama 10 dalių kryžminė validacija.

Rezultatai (1): mokymo imtis

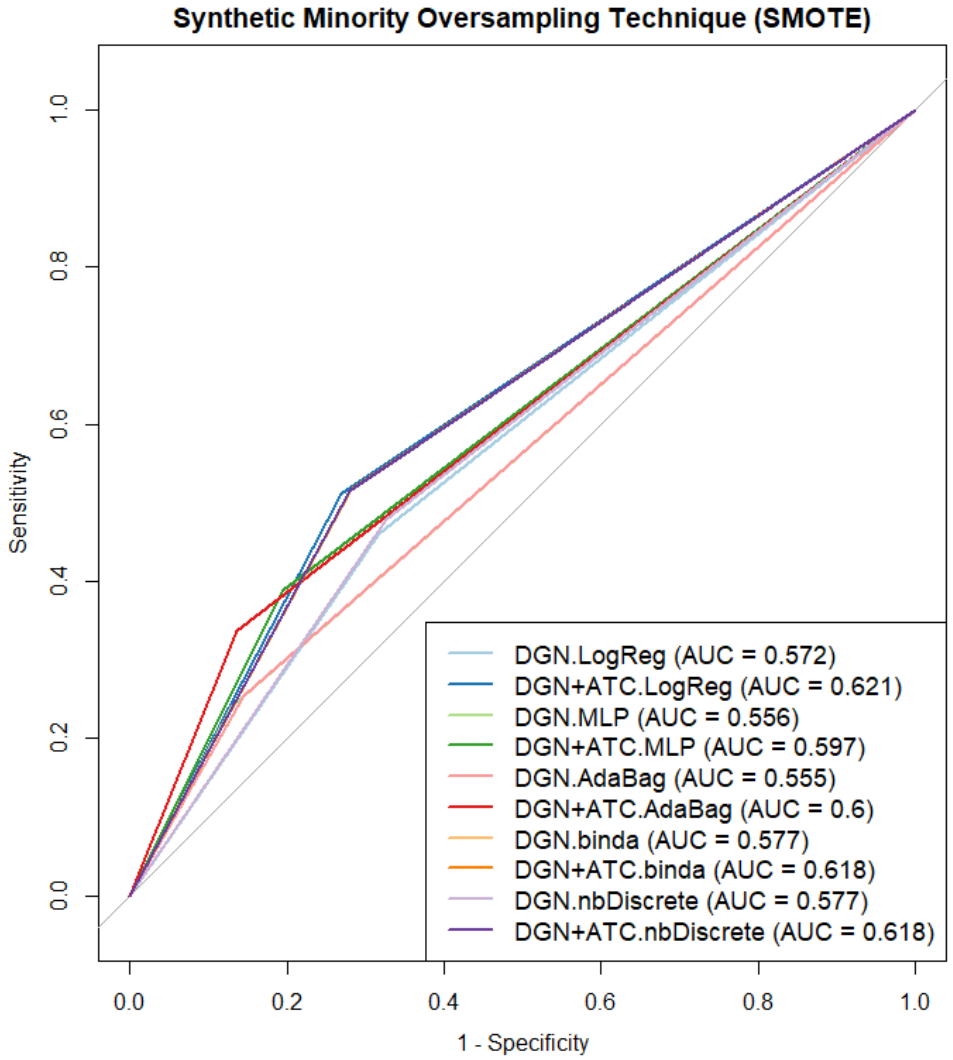


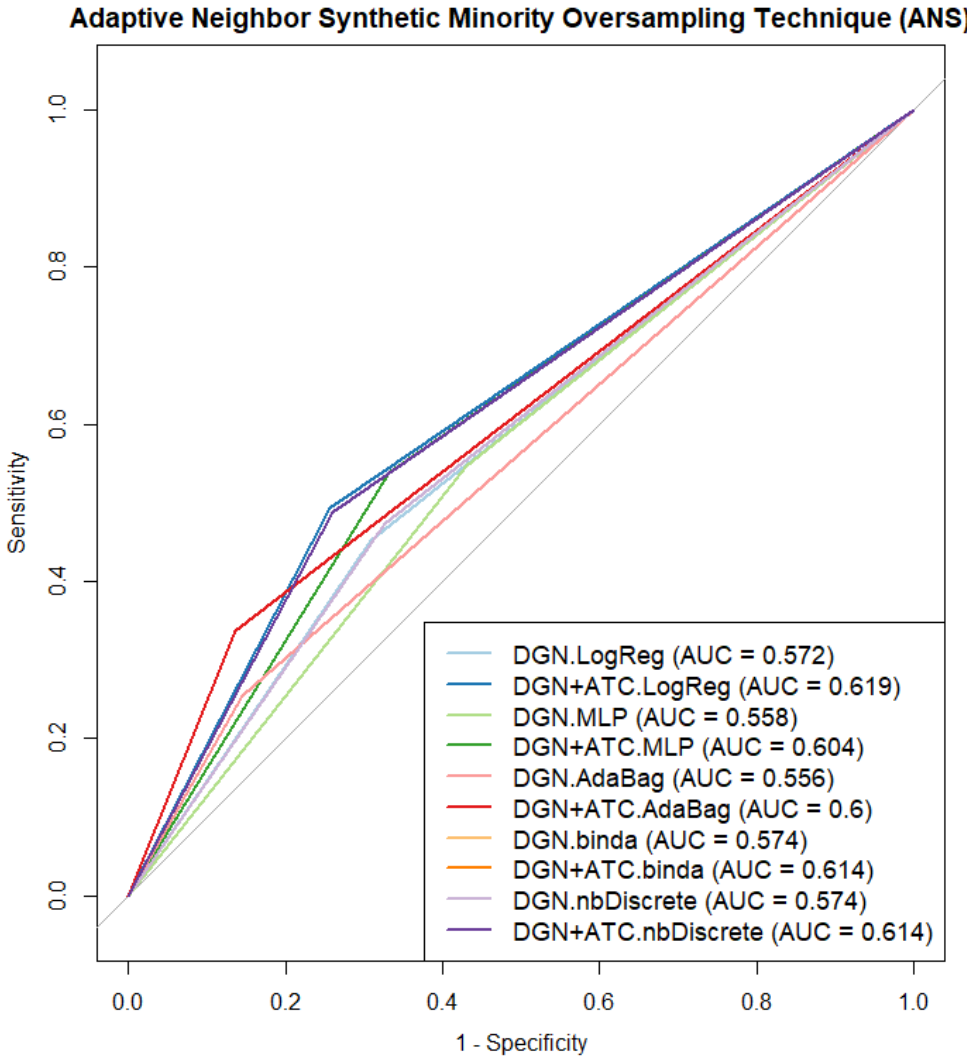


SMOTE
Aukščiausias AUC:
MLP, DGN+ATC,
AUC = 0,666

ANS
Aukščiausias AUC:
MLP, DGN+ATC,
AUC = 0,691

Rezultatai (2): testavimo imtis

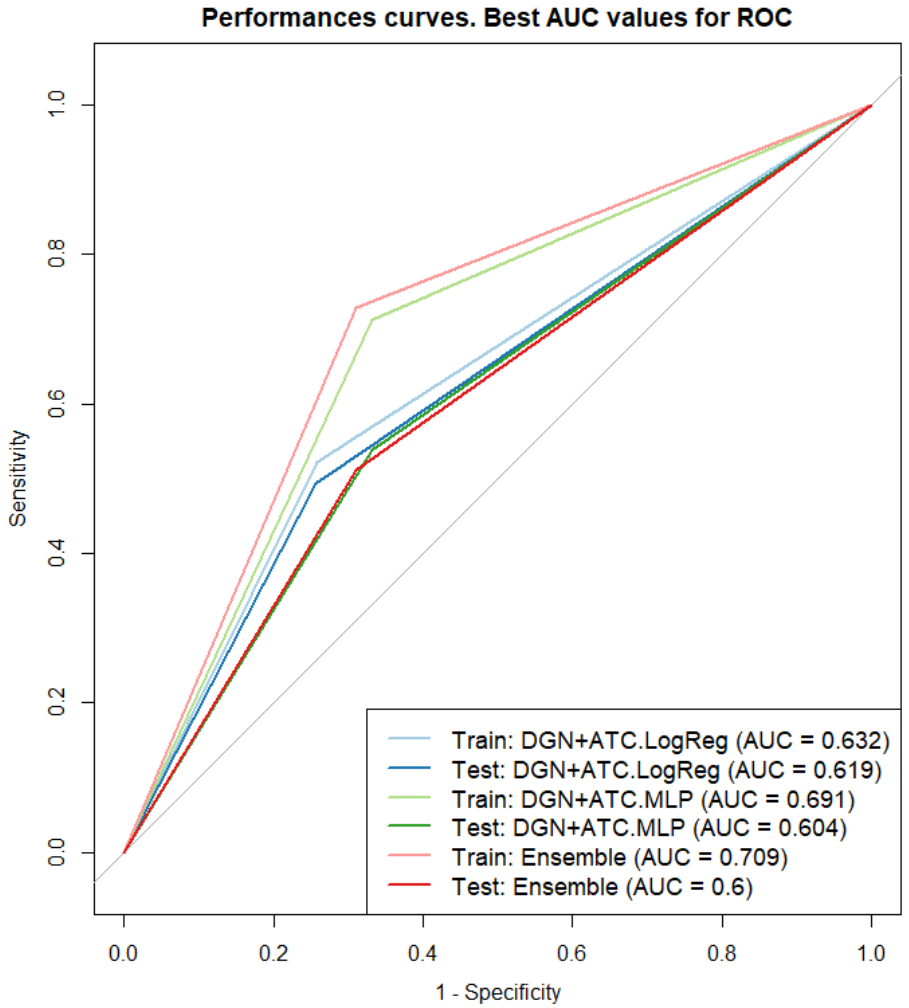


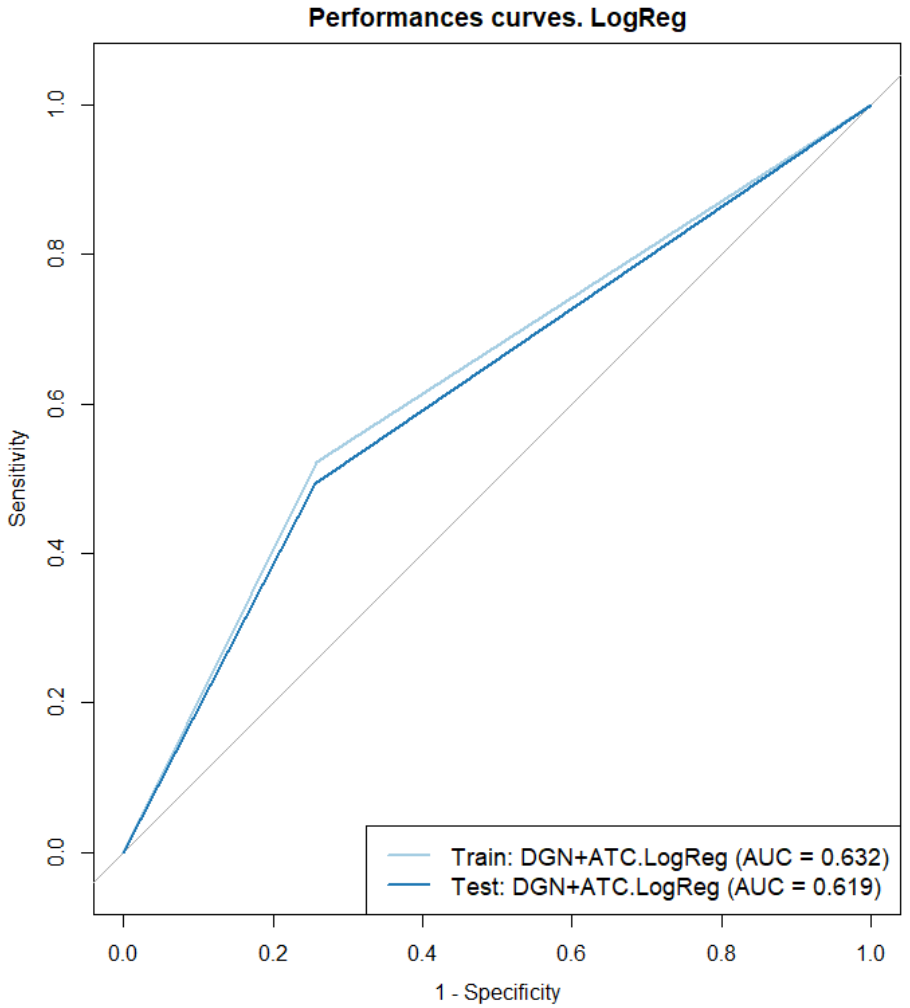


SMOTE
Aukščiausias AUC:
Log Reg, DGN+ATC
AUC = 0,621

ANS
Aukščiausias AUC:
Log Reg, DGN+ATC,
AUC = 0,619

Rezultatai (3): mokymo ir testavimo imtys





Iš literatūros apžvalgos:

- Vidutinis AUC 0,646
- Geriausias modelis: Log Reg, AUC = 0,782

Eksperimentai:

SMOTE
Aukščiausias AUC:
Log Reg, DGN+ATC
AUC = 0,621

ANS
Aukščiausias AUC:
Log Reg, DGN+ATC,
AUC = 0,619

Rezultatai (4): modelių tikslumo palyginimas

		DGN					DGN+ATC					
		LogReg	MLP	AdaBag	binda	nbDiscrete	LogReg	MLP	AdaBag	binda	nbDiscrete	Ensemble
SMOTE	Sensitivity	0,462	0,241	0,255	0,480	0,480	0,512	0,388	0,336	0,516	0,515	
	Specificity	0,682	0,872	0,856	0,675	0,675	0,730	0,805	0,864	0,719	0,720	
	AUC	0,572	0,556	0,555	0,577	0,577	0,621	0,597	0,600	0,618	0,618	
ANS	Sensitivity	0,452	0,547	0,255	0,474	0,474	0,494	0,539	0,337	0,488	0,488	0,511
	Specificity	0,691	0,569	0,856	0,674	0,675	0,744	0,668	0,864	0,741	0,741	0,690
	AUC	0,572	0,558	0,556	0,574	0,574	0,619	0,604	0,600	0,614	0,614	0,600

Apibendrinimas

Literatūros apžvalga

- Modeliai depresijai prognozuoti:
- Vidutinis AUC = 0,646, Geriausias modelis: Log Reg, AUC = 0,782

Eksperimentai

- Aukščiausias pasiektas tikslumas:
  - SMOTE Log Reg, AUC = 0,621 (jautrumas: 0,512, specifiškumas: 0,730)
  - ANS Log Reg, AUC = 0,619 (jautrumas: 0,494, specifiškumas: 0,744)
- Prediktoriai: Diagnozės + Vaistų terapinės grupės

Apibendrinimas

Probleminės vietos (literatūros apžvalga + eksperimentai)

- Didelis klasių disbalansas (~1 proc. depresijos)
- Ligos apibrėžimo kriterijai (klinikinės informacijos, objektyvaus įvertinimo trūkumas)
- Prediktorių parinkimas
- Mažas modelių jautrumas

Kito pusmečio planas

- Parengti publikaciją remiantis sisteminės literatūros apžvalgos duomenimis.
- Eksperimentai:
  - Kiti klasių subalansavimo metodai (pvz. didesnės sumažinimas)
  - Prognostinio modelio papildymas kaitos laike vertinimu.
  - Prognostinio modelio papildymas ligų/vaistų sąveikų informacija iš didelės apimties žinių duomenų bazių.