Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Įmonių bankroto procedūros pradžios prognozavimas

Baigiamasis magistro studijų projektas

Simonas Šilakauskas

Projekto autorius

Dr. Kęstutis Lukšys

Vadovas

Doc. dr. Aušrinė Lakštutienė

Vadovė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Įmonių bankroto procedūros pradžios prognozavimas

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Simonas Šilakauskas

Projekto autorius

Dr. Kęstutis Lukšys
Vadovas

Doc. dr. Aušrinė Lakštutienė
Vadovė

Doc. dr. Mindaugas Šnipas
Recenzentas

Doc. dr. Rasa Norvaišienė
Recenzentė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Simonas Šilakauskas

Įmonių bankroto procedūros pradžios prognozavimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Simono Šilakausko, baigiamasis projektas tema „Įmonių bankroto procedūros pradžios prognozavimas“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

Simonas Šilakauskas

(vardą ir pavardę įrašyti ranka)

(parašas)

Simonas Šilakauskas. Įmonių bankroto procedūros pradžios prognozavimas. Magistro studijų baigiamasis projektas, vadovai dr. Kęstutis Lukšys ir doc. dr. Aušrinė Lakštutienė; Kauno technologijos universitetas, matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika, Matematikos mokslai.

Reikšminiai žodžiai: *įmonių bankroto prognozavimas, klasifikavimas, disbalansas duomenyse, įmonių bankroto prognozei reikšmingi kintamieji.*

Kaunas, 2020. 54 p.

Santrauka

Iš anksto nustačius ar įmonei ateityje gali kilti finansinių sunkumų, galima anksti suteikti finansinę pagalbą arba užkirsti kelią turto išpardavimui ir pinigų iššvaistymui. Darbe nagrinėjama įmonių bankroto prognozavimo problema. Taikant atsitiktinių miškų mašininio mokymo algoritmą bei duomenų disbalanso kontrolę, nustatyta, kad modelis, apmokytas naudojant visų mėnesių duomenis, yra efektyvesnis nei atskiri modeliai, apmokyti specifinio mėnesio duomenimis, prognozuojant specifiniam mėnesiui į priekį. Taip pat, ištyrus disbalanso įtaką apmokymo duomenų imtyje, pastebėta, kad modeliai, apmokyti su didesniu disbalansu, geriau atskiria veikiančias įmones, o su mažesniu disbalansu – įmones, kurioms pradėta bankroto procedūra. Remiantis atsitiktinio miško rezultatais nustatyti reikšmingiausi kintamieji – nepriemoka VMI, nuosavas kapitalas, grynasis pelnas, darbuotojų skaičius įmonėje, įmonės amžius. Darbe taip pat pasiūloma klasifikavimo slenksčio vertė, priklausomai nuo disbalanso duomenyse bei klasifikavimo tikslo.

Simonas Šilakauskas. Forecast of bankruptcy procedure beginning for Lithuania's companies. Master's Final Degree Project, supervisors dr. Kęstutis Lukšys and assoc. prof. dr. Aušrinė Lakštutienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics, Mathematics Sciences.

Keywords: *forecast of bankruptcy, classification, data imbalance, significant variables for bankruptcy prediction.*

Kaunas, 2020. 54 p.

Summary

If one could assess whether a company will have financial difficulties in the near future, it would be possible to help the company in advance or to prevent owners from selling away assets and spending cash. The problem of company's bankruptcy forecasting is investigated in this paper. By applying random forest machine learning algorithm along with data imbalance controlling technique it was noticed that the model trained with all-months data is more efficient than the ones trained with specific month data, when forecasting one specific month ahead. Also, after investigating the impact of imbalance in the training data set, it was seen that the models, that were trained with higher imbalance, were better at distinguishing healthy companies, while models trained with smaller imbalance in the training data set were more efficient distinguishing companies that are going bankrupt. Based on random forest results, the most significant variables were determined – arrears to VMI, equity capital, net profit, total amount of employees, age of the company.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Nepriklausomai nuo teisinės ir ekonominės aplinkos, visada yra įmonių, kurios susiduria su finansiniais sunkumais ir nebūtinai su jais susitvarko, todėl bankrutuoja. Anksčiau nustačius įmones, kurioms ateityje gali kilti finansinių sunkumų, būtų galima iš anksto tam pasiruošti ir suteikti reikalingą pagalbą, taip išvengiant papildomų rūpesčių.

Tyrimo aktualumas. Bankrutuojančios įmonės neretai gali sukelti nemalonumų ne tik su įmone susijusiems asmenims, bet ir kreditoriams. Anksčiau nustačius įmones, kurioms gali būti pradėta bankroto procedūra, galima užkirsti kelią įmonės turto išsipardavimui ir pinigų iššvaistymui. Darbe siekiama nustatyti, kaip kuo tikslingiau panaudoti turimus duomenis bankroto procedūros pradžios prognozavimui.

Tyrimo problematika. Efektyvus turimų duomenų panaudojimas bankroto procedūros pradžios prognozavimui. Bankroto procedūros pradžios prognozavimas.

Tyrimo naujumas. Išanalizuotoje literatūroje daugiausiai naudojami metiniai įmonės finansinių rodiklių duomenys. Šiame tyrime įtraukiami ir mėnesiniai duomenys apie darbuotojus bei ištiriama įvairių duomenų aspektų įtaka bankroto prognozei.

Tyrimo tikslas. Sudaryti įmonių bankroto procedūros iniciavimo trumpalaikės prognozės modelį, naudojant metinius ir mėnesinius duomenis.

Tyrimo uždaviniai:

1. Atskleisti bankroto prognozavimo modelių taikymo problematiką ir nustatyti pagrindinius žingsnius ir metodus, sudarant bankroto prognozės modelius;
2. Sutvarkyti ir tinkamai paruošti gautus duomenis mašininio mokymosi algoritmams;
3. Nustatyti ar skirtingi įmonių finansinių rodiklių vėlavimo periodai daro skirtingą įtaką bankrotui;
4. Ištirti disbalanso įtaką apmokymo duomenų imtyje bankroto procedūros pradžios prognozei;
5. Pateikti rekomendacijas klasifikavimo tikimybės slenksčio nustatymui;
6. Nustatyti bankroto prognozei reikšmingiausius kintamuosius.

1. Literatūros analizė

Bankrotas yra plačiai nagrinėjama problema ekonomikos mokslo srityje. Šis procesas turi didelę įtaką ne tik pačiam verslui, bet ir jo aplinkai – vartotojams, investuotojams [23].

Klasifikuojant retus įvykius, šiuo atveju – įmonės bankrotą, susiduriama su įvairiais uždaviniais: klasių disbalansas, reikšmingų ir svarbiausių kintamųjų atsirinkimas, klasifikavimo metrikų pasirinkimas, šalies, regiono bei sektoriaus niuansai. Daugumoje atliktų tyrimų, susijusių su klasifikavimo problema, šie minėti faktoriai išskiriami, kaip svarbiausi tyrimo eigai, kadangi nuo jų priklauso klasifikavimo tikslumas, algoritmų greitaveika bei jų efektyvumas.

1.1. Bankrotą sąlygojantys veiksniai ir priežastys

Nagrinėjant bankroto priežastis, tyrėjai į tai žvelgia iš įvairių pusių. Venkataramana [24] ir Ahmedas Khaliq‘as [25] teigia, kad įmonės bankrotas yra situacija, kai finansiniai įsipareigojimai viršija turimą turtą, o to priežastys yra neefektyvus turimų resursų panaudojimas, nepakankamai užtikrintas įmonės veiklų valdymas, pardavimų kritimai ir bendras rinkos padėties blogėjimas. Oliveris Lukasonas ir Ričardas Hoffmanas [26] papildo Venkataramaną ir į bankrotą žvelgia kaip į paskutinę įmonės ekonominio nuosmukio būseną. Tyrėjai pabrėžia, kad bankrotas yra ne momentinis, o laike besitęsiantis procesas, kurio trukmė labai įvairiai svyruoja.

Bankroto priežastis detaliau nagrinėjo tyrėjai Hubertas Ooghe‘us ir Sofija De Prijcker [17]. Remdamiesi 12 Belgijos įmonių istorija, savo tyrime jie išskyrė keturis įmonių žlugimo proceso tipus, kuriuos apžvelgiant detaliau buvo išskirti bendri faktoriai, darantys įtaką įmonių bankrotui.

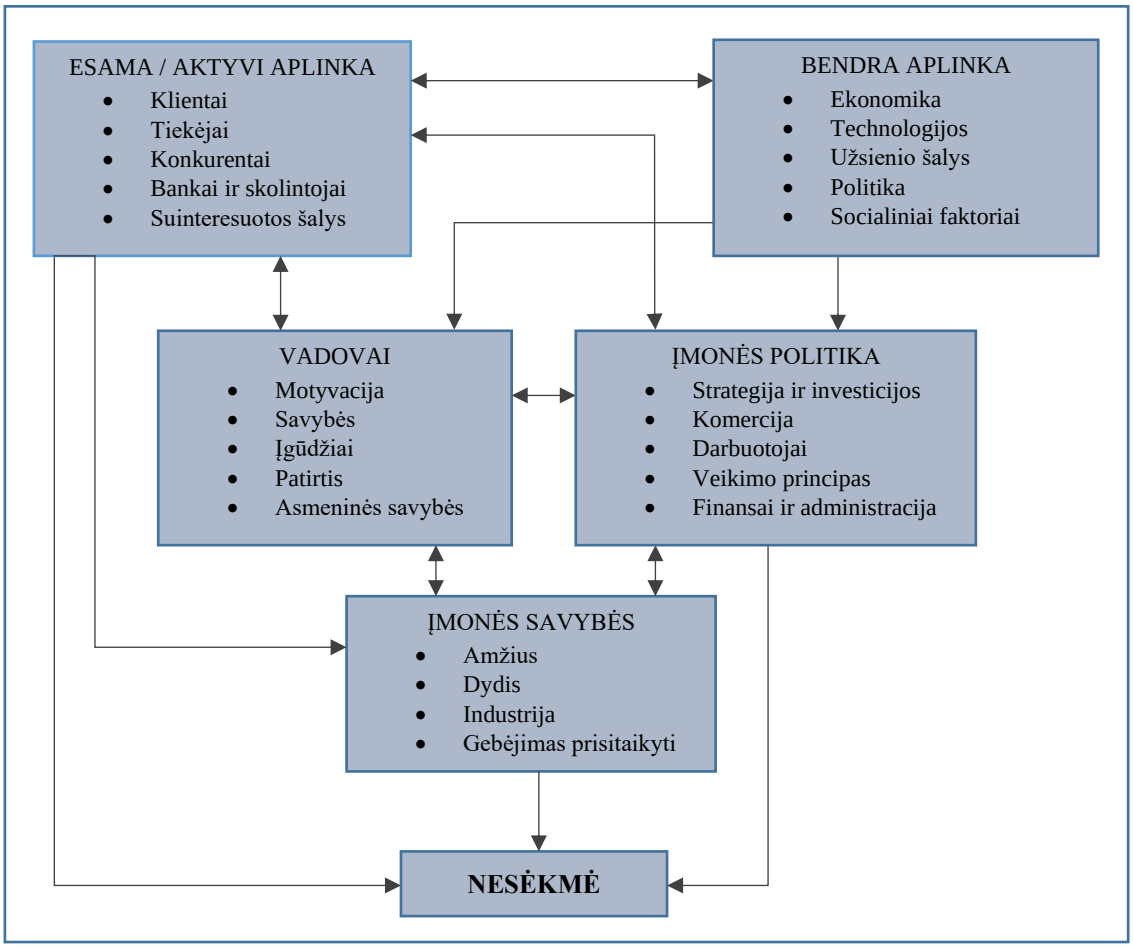
1) Iš visų nagrinėtų įmonių, pirmas išskirtas tipas buvo naujos įmonės arba „startuoliai”. Apžvelgus tokių įmonių istoriją buvo pastebėta, kad jose vadovams stipriai trūkdavo valdymo patirties bei žinių, susijusių su pačia įmonės industrija. Šios kompanijos neturėjo strateginio pranašumo prieš konkurentes, pasižymėjo prasta vidine organizacine sistema, turėdavo finansinių problemų nuo pat įmonės įkūrimo. Išoriniai faktoriai tokių įmonių bankrotui reikšmės neturėdavo.

2) Antras išskirtas tipas buvo įmonės, kurių patyrę, rizikuoti linkę vadovai vadovaudavosi ambicingais planais plėsti įmonės veiklą, tačiau dėl klaidingos informacijos ir pervertintos strategijos, apyvartos planų, dažnai nukentėdavo. Vėliau pinigai skiriami atkūrimo strategijoms pradėdavo varžyti įmonės veiklą ir ji tapdavo mažiau prisitaikanti prie aplinkos veiksnių. Nors restruktūrizavimo planai padėdavo įmonėms iš lėto atsigauti po nesėkmingų investicijų, tačiau staigesni pokyčiai rinkoje privesdavo įmones prie bankroto.

3) Trečiasis tipas buvo išskirtas įmonių, kurių vadovai priimdavo nepasvertas rizikas, neatsižvelgdami į organizacijos struktūrą. Šios įmonės greitai nusilpdavo finansiškai bei prarasdavo aplinkos pasitikėjimą.

4) Ketvirtajam tipui buvo priskirtos įmonės, kurios veikė sėkmingai ir generavo pelną, tačiau jų vadovai nebuvo pakankamai ryžtingi ir motyvuoti. Tokios įmonės mažai reaguodavo į aplinkos veiksnius, pardavimų kritimai būdavo pastebimi per vėlai, tai susilaukdavo suinteresuotų šalių nepasitikėjimo ir ilgainiui sukeldavo finansinį nuosmukį

Apibendrinantys veiksniai pateikti 1 pav.



1 pav. Pagrindiniai įmonių bankroto faktoriai ir veiksniai (sudaryta autoriaus pagal [17])

Išanalizavę visus šiuos įmonių žlugimo tipus, autoriai išskyrė kelis pagrindinius veiksnius – bendrus arba turinčius sąryšių nagrinėtiems tipams: įmonės valdytojų klaidos; neatitikimai įmonės politikoje ir naudojamoje strategijoje; išorinės bei vidinės aplinkų veiksniai.

1.2. Bankrotas Lietuvoje

Lietuvos Respublikos finansų ministerijos bei Lietuvos statistikos departamento duomenimis 2014-2017 metais pradėtų bankroto procedūrų skaičius Lietuvoje augo ir 2016 - 2017 metais buvo pradėta daugiausiai bankroto procedūrų: 2741 ir 2978. Kritimas pastebimas 2018 metais – buvo pradėta 2094 bankroto procesų [35]. Bankroto procedūrų augimas aiškinamas 2015 metais pakitusia bankroto iniciavimo procedūra – bankroto iniciavimo procesas buvo supaprastintas ilgai nebeveikiančioms, daug skolų sukaupusioms įmonėms. Taip pat dėl bankroto proceso skelbimo supaprastinimo didžioji dalis bankrutuojančių įmonių buvo 5-10 ar 10 ir daugiau metų gyvavusios įmonės [36]. Visų 2017 metais pradėtų bankroto procesų didžioji dalis (beveik 96%) įmonių buvo mažos, jose dirbo iki 10 darbuotojų.

1.3. Kintamųjų, naudojamų bankrotui prognozuoti, tipai ir kategorijos

Siekiant efektyviai prognozuoti bankrotą, vienas svarbiausių uždavinių yra teisingų kintamųjų (finansinių rodiklių) atrinkimas. Pasak Philippo du Jardino [10] net ir turint daug informacijos, kurią aprašo daug kintamųjų, kuriant modelį reikia parinkti kuo mažiau kintamųjų, kurie duoda kuo daugiau naudingos informacijos. Tyrėjas savo tyrime išskiria tris kategorijas kintamųjų, kurie geriausiai atspindi firmos nesėkmę:

1. firmos finansinių ir kitų pagrindinių charakteristikų kintamieji: balanso lapas, pajamų deklaracijos kintamieji, struktūra, valdymas, produkciją aprašantys kintamieji;
2. firmos aplinkos kintamieji: palūkanų norma, augimas, indikatoriai susiję su sektoriumi;
3. kintamieji susiję su finansų rinkomis ir kaip rinka vertina kompaniją – akcijos kaina.

Straipsnio autorius išskiria kelis pagrindinius būdus, kaip yra konstruojami minėtų kategorijų kintamieji pagal populiarumą:

- finansiniai santykiniai rodikliai – dviejų rodiklių santykis;
- statistiniai kintamieji – vidurkiai, variacija, įvairūs įverčiai;
- variacijos kintamieji – santykinio arba vieno finansinio kintamojo evoliucija laike;
- nefinansiniai kintamieji – kitos kompanijos ar jos aplinkos charakteristikos;
- rinkos kintamieji – santykiai arba kintamieji, susiję su akcijų kaina ar jų grąža.

Phillipas [10] savo apžvalgoje taip pat nustatė, kokie metodai yra dažniausiai naudojami bankroto prognozavimui kintamiesiems atrinkti:

- populiarumas literatūroje ir ankstesniuose tyrimuose;
- vienanarė analizė – t kriterijus, F kriterijus, koreliacijos kriterijai;
- pažingsninė paieška – *Wilko lambda*, tikėtinumo kriterijus;
- genetiniai, specialūs algoritmai – *Relief, Tabu*;
- ekspertų įžvalgos ir nuomonė;
- metodai, su kuriai pritaikomos ne tiesinės modeliavimo technikos – neuroniniai tinklai;
- kita – regresinė analizė, regresijos medžiai, teoriniai modeliai;

1.3.1. Finansinių rodiklių taikymas tyrimuose

Anksčiau apžvalgoje išskirtų kintamųjų kategorijų sąveiką prognozės modeliuose bei kintamųjų atrinkimą, priklausomai nuo laikotarpio horizonto (laiko iki bankroto) tyrė Amerikos ir Kinijos mokslininkai Shaonanas Tianas, Yanas Yu ir kt. [8]. Tikslas buvo nustatyti ar rinkos rodikliai ir įmonių finansiniai santykiniai rodikliai gali būti naudojami kartu efektyviai prognozuoti bankrotą bei įvertinti, kokią įtaką laikotarpio horizontas turi reikšmingų kintamųjų atrinkimui. Tyrimo duomenų imtį sudarė 17570 Amerikos įmonių duomenys nuo 1980 iki 2009 metų. Iš pirminių 39 kintamųjų buvo atrinkti 7 įmonės (žr. 1 lentelė) ir 5 rinkos (žr. 2 lentelė) finansiniai santykiniai rodikliai.

1 lentelė. X pažymėti atrinkti įmonės rodikliai [8].

Kintamieji \ Laikas iki bankroto	**1 mėn.**	**6 mėn.**	**12 mėn.**	**24 mėn.**	**36 mėn.**	**60 mėn.**
Visos skolos / visas turimas turtas	X	X	X	X	X	X
Turimi įsipareigojimai / visas turimas turtas	X	X	X	X	X	X
Turimi įsipareigojimai / pardavimai						X
Visas turimas turtas					X	X
Veiklos pajamos / visas turimas turtas					X	X
Nuosavas kapitalas / visas turimas turtas					X	

Nagrinėjant įmonės rodiklius, esant trumpesniam laikui iki bankroto – nuo 1 mėnesio iki 2 metų, buvo daugiausiai naudojami santykiniai rodikliai sudaryti iš skolų, įsipareigojimų ir turto. Prognozės laikotarpiui esant nuo 3 iki 5 metų reikšmingų kintamųjų padaugėjo ir šiuo atveju svarbūs tapo pardavimai, pardavimai, pajamos ir nuosavas kapitalas.

2 lentelė. X pažymėti atrinkti rinkos rodikliai [8].

Kintamieji \ Laikas iki bankroto	**1 mėn.**	**6 mėn.**	**12 mėn.**	**24 mėn.**	**36 mėn.**	**60 mėn.**
Visi įsipareigojimai / (rinkos kapitalas + visi įsipareigojimai)	X	X	X	X		X
Grynasis pelnas / (rinkos kapitalas + visi įsipareigojimai)	X	X	X	X		
Akcijos kaina	X	X	X	X	X	X
Akcijos kintamumas	X	X	X	X	X	
Grąža lyginant su S&P 500 indeksu	X	X	X	X		

Nagrinėjant rinkos rodiklius, nuo 1 mėnesio iki 2 metų visi rodikliai buvo svarbūs, prognozuojant nuo 3 iki 5 metų beveik pusė rodiklių praranda svarbą ir yra nebenaudojami. Svarbiausia per visus laikotarpius išlieka akcijos kaina.

Dalį mokslininkų atrinktų ir naudotų kintamųjų [8], taip pat sėkmingai pritaikė ir mokslininkai iš Belgijos finansų ir apskaitos departamento – L. Cultrera bei X. Bredart‘as [13]. Jie atliko tyrimą, kurio metu siekė sukurti modelį mažų ir vidutinio dydžio Belgijos įmonių bankrotui prognozuoti. Tyrimui buvo naudoti duomenys apie 7152 Belgijos įmones, iš kurių 3576 buvo bankrutavusios tarp 2002 ir 2012 metų. Tikslas buvo sukurti gerą įmonių bankroto prognozės modelį, kuris remtųsi įmonių finansine struktūra, pelningumu, likvidumu ir mokumu. Tyrimui pasirinkti nepriklausomi kintamieji: turimo turto vertės ir finansinių įsipareigojimų santykis, pajamų prieš mokesčius ir turimo turto vertės santykis, turimo laisvojo kapitalo ir turimo turto vertės santykis, išlaidų mokesčiams ir sukurtos vertės santykis, pinigų srauto ir visos skolos santykis. Buvo pastebėta, kad labiau linkusios bankrutuoti buvo jaunesnės ir mažesnės įmonės bei įmonės, užsiimančios statybų, maitinimo, žemdirbystės ir industrijų sektorių veikla.

Bagheras [20], tirdamas Kipro akcijų biržos įmones taip pat kaip ir Cultrera [13] naudojosi kintamaisiais, kurie buvo išskirti kitoje įvairioje literatūroje bei pagrindinį dėmesį skyrė įmonės finansiniams rodikliams. Apskaitos rodikliai šiame tyrime yra traktuojami kaip vieni svarbiausių, kadangi jie pakankamai gerai charakterizuoja įmonę, tačiau autorių nuomone, norint tiksliai prognozuoti įmonės bankrotą reikia svarstyti rinkos bei makroekonomikos aspektus – panašios

struktūros tyrimas buvo apžvelgtas Amerikos bei Kinijos mokslininkų straipsnyje [8]. Baghero straipsnyje buvo sudaryta tokia kintamųjų schema:

3 lentelė. Modelio sudarymui naudojami kintamieji ir santykiniai rodikliai [20]

Aspektas	**Kintamasis**		**Apskaičiavimas**
Apskaita	Pelningumo koeficientas	Turto pelningumo rodiklis	$\frac{grynasis\ pelnas}{akcijų\ vertė}$
		Pastovaus kapitalo pelningumo rodiklis	$\frac{pelnas\ prieš\ palūkanas - mokesčiai(EBIT)}{darbinis\ kapitalas}$
	Svertinis koeficientas	Skolos santykis	$\frac{visi\ įsipareigojimai}{visas\ turtas}$
	Likvidumo koeficientas	Apyvartinio kapitalo ir turto santykis	$\frac{apyvartinis\ kapitalas}{visas\ turtas}$
	Aktyvumo koeficientas	Turto apyvartumo rodiklis	$\frac{pardavimų\ pajamos}{visas\ turtas}$
		Atsargų apyvartumo rodiklis	$\frac{pardavimų\ savikaina}{atsargos}$
Rinka	Įmonės bendra vertė su visa skola		$\frac{rinkos\ vertė}{visa\ skola}$
	Akcijų kaina		$log(akcijų\ kaina)$
	Akcijų grąža		$\frac{(akc.kain._t - akc.kain._{t-1}) + pelnas\ vienai\ akcijai}{akc.kain._{t-1}}$
Makroekonomika	Infliacijos lygis		$vartotojų\ kainų\ indeksas\ (CPI)$
	Palūkanų norma		$palūkanų\ dorma\ indėliams\ iš\ gyventojų\ (iki\ 3\ mėn.)$
	Ekonomikos augimas		$\frac{BVP_t - BVP_{t-1}}{BVP_{t-1}}$

Buvo gauta, kad daugiausiai sąryšio su įmonės bankrotu turi finansiniai apskaitos ir rinkos rodikliai. Didžiausią įtaką modelyje darantys rodikliai buvo: investicijų grąža (angl. *capital employed*), turto pelningumo rodiklis (angl. *return on assets*), skolos koeficientas (angl. *debt ratio*), apyvartinio kapitalo ir turto santykis (angl. *working capital to assets ratio*), turto apyvartumo rodiklis (angl. *asset turnover ratio*), akcijų kaina (angl. *stock price*) ir rinkos kapitalizacija (angl. *market capitalization*). Investuotojams Kipro akcijos biržoje buvo rekomenduojama vadovautis šiais rodikliais. Prieštaringai nei Shaonan‘o tyrime [8], šio straipsnio [20] autoriai rekomenduoja bankroto prognozės modelių sudarymui nenaudoti makroekonomikos rodiklių, kadangi jie beveik neteikia naudingos informacijos apie įmonės bankroto riziką.

Dalis Baghero [20] išskirtų kintamųjų taip pat yra naudojami tiriant Slovakijos kompanijų duomenis. Tyrėjos iš Slovakijos Ivana Podhorska, Maria Kovacova ir Katarina Valaskova ištyrė daugiau nei 8000 finansinių 2016 metų įrašų iš Slovakijos kompanijų [21]. Tyrimo tikslas buvo nustatyti statistiškai reikšmingus rodiklius, kurie koreliuoja (remiantis Pirsono koreliacijos koeficientu) su įmonės kapitalo – skolos santykiu, kuris buvo pasirinktas kaip įmones klasifikuojantis kintamasis. Buvo nustatyta, kad su klasės kintamuoju (įmonės kapitalo ir skolos santykiu) vidutiniškai koreliuoja 3 finansiniai rodikliai: turto pelningumo rodiklis, turimo turto ir trumpalaikių įsipareigojimų santykis bei įmonės laisvojo turto ir trumpalaikių skolų santykis.

Galima pastebėti, kad dalis išskirtų finansinių rodiklių sutampa su anksčiau apžvelgtame straipsnyje išskirtais pagrindiniais rodikliais.

Tiriant, kaip skiriasi bankroto prognozės kintamieji pagal regioną, Japonijos tyrėjai Mingas Xu ir Chu Zhangas [11] siekė išsiaiškinti ar kintamieji, naudojami Amerikos rinkos bankroto prognozės modeliuose [8], gali būti pritaikomi vietinei Japonijos rinkai. Buvo padaryta išvada, kad ne visi Amerikos rinkai aktualūs kintamieji buvo reikšmingi šiam tyrimui, tačiau pridėjus papildomus kintamuosius, kurie įvertina tiriamos šalies rinkos specifiką ir sujungus egzistuojančius prognozės modelius galima pasiekti gerą klasifikavimo tikslumą.

Taip pat tyrėjai iš Japonijos savo sukurtu modeliu vertino įmonių ryšio stiprumo su bankų grupe poveikį įmonių bankrotui. Buvo padaryta papildoma išvada, kad įmonės turinčios stipresnius ryšius su bankais ar jų grupėmis yra linkusios bankrutuoti rečiau.

Tiriamo regiono specifikas, prognozuojant bankrotą taip pat nagrinėjo Davidas Alaminos‘as ir kt. [19]. Tyrėjų tikslas buvo ištirti ar įmanoma sukurti ir sėkmingai pritaikyti globalų įmonių bankroto prognozavimo modelį ir nustatyti, kaip skiriasi esminiai kintamieji, prognozuojant bankrotą. Tyrimui buvo naudoti 440 įmonių, priklausančių įvairiems regionams (Azijos, Europos ir Amerikos) duomenys, iš daugiau nei 20 šalių nuo 1990 iki 2013 metų. Taikant modelį buvo remiamasi pagrindiniais finansiniais rodikliais, kurie buvo naudojami kituose autorių apžvelgtuose straipsniuose. Pasirinkti rodikliai apibendrinantys tokius aspektus: pelningumas, įsiskolinimas, likvidumas ir įmonės efektyvumas. Taip pat pridėti fiktyvūs kintamieji, nusakantys, kokiam regionui ir kokiai industrijai priklauso įmonė (*GICS – Global Industry Classification Standart*). Suklasifikavus duomenis pagal kiekvieną regioną ir globaliai, buvo gauta, kad skirtingiems regionams didžiausią įtaką darė skirtingi faktoriai: Azija – pelningumas ir įsiskolinimas; Europa – pelningumas, įsiskolinimas ir likvidumas; Amerika – apyvartinis kapitalas, pelnas prieš palūkanas ir mokesčius, pardavimai ir įsiskolinimai.

Kitaip nei dauguma mokslininkų, amerikiečių tyrėjai Shyamas B. Bhandaris ir Rajesh‘as Iyeris [16] tyrė Amerikos įmonių bankrotą remdamiesi kintamaisiais, paremtais įmonių pinigų srautu (angl. OCF - *operating cash flow*). Buvo dirbama su 100 įmonių duomenimis iš 2008 - 2010 metų laikotarpio. Tyrimui pasirinkti kintamieji: pinigų srauto ir turimų įsipareigojimų; pinigų srauto, palūkanų ir mokesčių santykis su palūkanomis; pinigų srauto ir pardavimų santykis; pinigų srauto ir turimo turto santykis; pajamų kokybė; kritinis likvidumo koeficientas (angl. *quick ratio, acid test*) ir trejų metų pardavimų augimai.

4 lentelė. Finansinių rodiklių literatūros analizės apibendrinimas (sudaryta autoriaus)

Straipsnis \ kintamieji	Visas turtas	Trump. įsipareig	Visi įsipareig	Pard. pajamos	Apyvart. kapit.	Peln. prieš apm.	Trump. turt.	Grynieji	Atsargos
E. Fedorova ir E. V. Gilenko (2013) [5]	X	X	X	X				X	X
S. Tian'as, Y. Yu ir kt. (2015) [8]	X	X	X	X				X	X
M. Xu ir C. Zhang'as (2009) [11]	X	X	X		X				
K. Pavol'as, M. Fleischer'ė ir kt. 2016) [12]	X			X	X	X			
L. Cultrera ir X. Bredart'as (2016) [13]	X	X				X	X		
N. B. Misu ir E. S. Codreanu (2014) [14]	X	X	X	X		X	X	X	X
M. Celli (2015) [15]	X			X	X	X			
B. Shyam'as ir I. Rajesh'as (2013) [16]	X	X		X	X	X		X	X
D. Alaminos'as, A. Castill'as ir kt. (2016) [19]	X	X		X	X	X	X		
B. Asgarnezhad'as ir M. Soltani (2016) [20]	X	X	X	X		X	X	X	X

Populiariausi literatūroje finansiniai rodikliai pateikti 4 lentelėje. Galima pastebėti, kad dauguma tyrėjų didžiausią dėmesį kreipia į įmonių finansinius rodiklius – jų santykius. Svarbiausi pavieniai rodikliai – visas turtas, trumpalaikiai įsipareigojimai, pajamos. Mažiau naudojami yra rinkos kintamieji bei makroekonomikos kintamieji.

1.3.2. Finansinių rodiklių atrinkimo metodų analizė

Nors dauguma mokslininkų, atlikdami savo tyrimus [11, 13, 19, 20] ir rinkdamiesi kintamuosius, remdavosi ankstesniais tyrimais, dalis tyrėjų reikšmingų kitamųjų atrinkimo procesą nagrinėjo detaliau ir tam pasitelkdavo įvairius metodus.

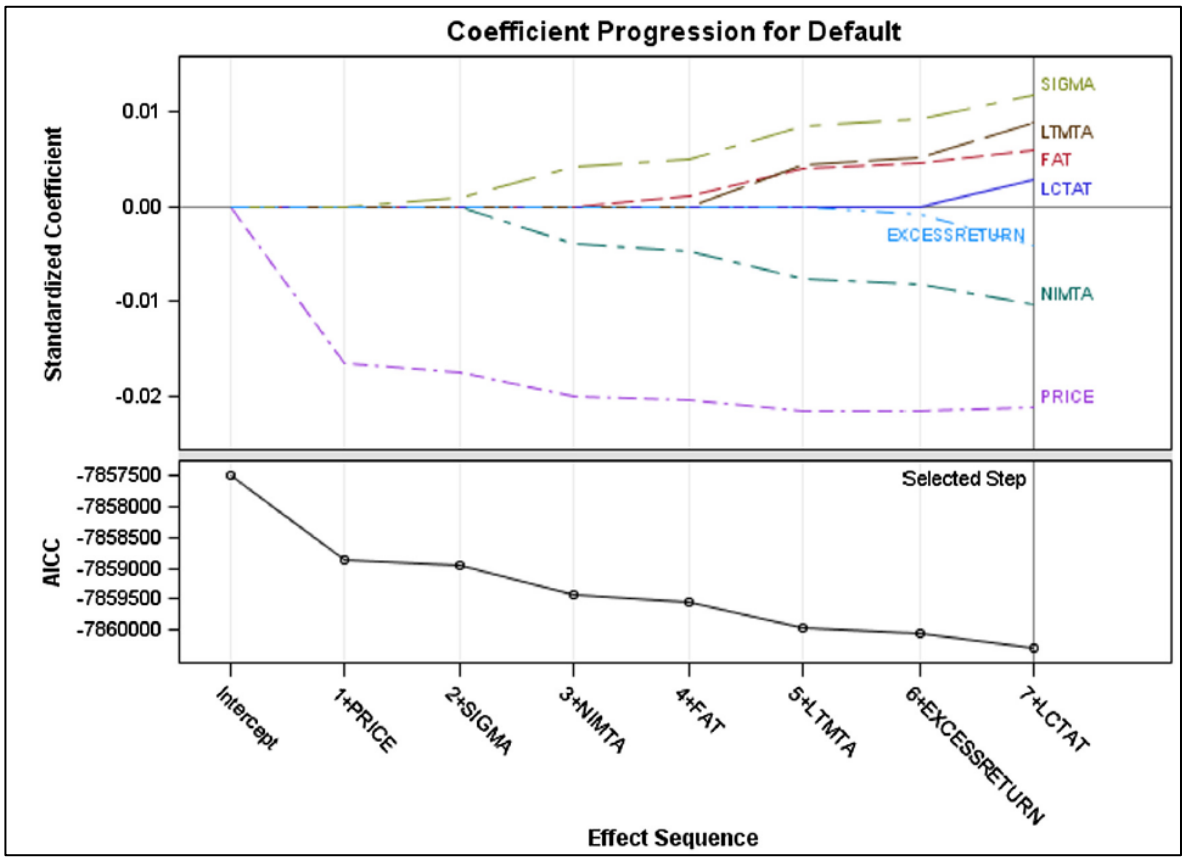
Tyrėjas iš Taivano Chih'as-Fong'as Tsai'us savo tyrime lygino 5 reikšmingų kintamųjų atrinkimo metodus: t-kriterijus, koreliacijų matrica, pažingsninė regresija, pagrindinių komponenčių analizė ir faktorinė analizė [9]. Klasifikavimui ir prognozių pagal atrinktus kintamuosius tikslumui įvertinti buvo naudojami neuroniniai tinklai. Kintamųjų atrinkimas ir klasifikavimas buvo naudojamas keliems skirtingiems bankroto duomenų rinkiniams. Priklausomai nuo duomenų rinkinio, kintamųjų skaičius buvo nuo 14 iki 39 (kintamieji straipsnyje nenurodyti), o įmonių - nuo 240 iki 2528. Gauti rezultatai pateikiami 5 lentelėje.

5 lentelė. Klasifikavimo tikslumas ir atrinkų kintamųjų skaičius, priklausomai nuo kintamųjų atrinkimo metodo - stulpelyje kint. nurodyta, kokia dalis visų kintamųjų buvo palikta (sudaryta autoriaus pagal [9])

	T-kriterijus		**Pažingsninis**		**Koreliacijų matrica**		**Faktorinė analizė**		**Pagrindinių komp. analizė**	
	Kint.	Tiksl.	Kint.	Tiksl.	Kint.	Tiksl.	Kint.	Tiksl.	Kint.	Tiksl.
Japonijos	80%	63.53	33.3%	82.64	80%	60.16	73.3%	74.22	54.3%	74.00
Australijos	85.7%	89.27	50%	84.74	85.7%	89.31	64.3%	86.08	64.3%	89.93
Vokietijos	60%	75.87	50%	75.51	60%	74.84	80%	78.76	45%	67.03
UC	69.2%	97.25	33.3%	96.33	69.2%	96.70	82.1%	97.30	41%	96.47

Buvo gauta, kad vidutiniškai T-kriterijaus atrinkimo metodas buvo geriausias lyginant su kitais tikslumo atžvilgiu, stabilumu ir paklaidų dydžiu. Daugiausiai kintamųjų pašalino pažingsninis kintamųjų atrinkimo metodas, juo atrinktų kintamųjų modelio klasifikavimo tikslumas stipriai neatsiliko nuo kintamųjų, atrinktų naudojant T-kriterijaus metodą.

Anksčiau apžvelgtame Kinijos ir Amerikos mokslininkų straipsnyje [8] kintamųjų atrinkimui buvo taikomas Lasso metodas. Atrinktų kintamųjų vertinimui buvo naudojamas AICC kriterijus. Iš pradinių 39 kintamųjų atrinkti 16, iš kurių vienu metu modelyje buvo taikomi 7, priklausomai nuo trukmės iki bankroto (žr. 2 pav.).



2 pav. Lasso metodu pridedami kintamieji. Atranka vykdoma pagal AICC kriterijų [8].

Elena Fedorova, Evgeni Gilenko ir kt. [5] prognozei reikšmingų kintamųjų atrinkimui taikė dispersinę analizę ir klasifikavimo algoritmus: daugialypė diskriminantinė analizė (angl. MDA - *multivariate discriminant analysis*), klasifikavimo ir regresijos medis (angl. CRT - *classification*

and regression tree) ir logistinė regresija. Taip pat dvi kintamųjų grupės buvo sudarytos remiantis Rusijos teisės aktų rekomenduojamais kriterijais „118-MinEcon“ ir „367-GovRF“. Geriausias klasifikavimo modelis buvo gautas kintamuosius atrinkus taikant logistinę regresiją. Taip pat buvo bandoma juos jungti: buvo renkami tie kintamieji, kurie buvo atrinkti bent su 2 iš 3 taikytų algoritmų – taip iš viso gaunant 6 reikšmingus kintamuosius. Šiuo atveju buvo gautas net geresnis klasifikavimo tikslumas. Straipsnio autoriai pabrėžė reikšmingų kintamųjų atrinkimo svarbą, siekiant prognozuoti įmonės bankrotą.

Iš dalies panašiu principu kaip ir Rusijos tyrėjai [5], reikšmingų kintamųjų procesą atliko ir Amerikos mokslininkai Sheikhas Rabiul‘as, Islamas Williamas Eberle ir kt. [22] kombinavo įvairių algoritmų veikimą, tačiau patį kintamųjų atrinkimo procesą atliko etapais ir aprašė jį kiek išsamiau. Tyrėjai šį procesą suskirstė į tris žingsnius:

1. Pašalino visus su laiku susijusius kintamuosius. Autoriai šį pasirinkimą grindė tuo, kad savo tyrime jie nedaro jokios laiko eilučių analizės, todėl laiką nusakantys kintamieji yra pertekliniai ir nereikalingi;
2. Remiantis koreliacijos koeficientu, buvo pašalinti kintamieji, kurie buvo stipriai tarpusavyje susiję. Kintamieji, nešantys tą pačią ar labai panašią informaciją yra pertekliniai.
3. Atsitiktinio miško ir genetinio algoritmų pagalba buvo pasirinktas galutinis kintamųjų rinkinys.

Taip 2 ir 3 žingsnių metu iš kintamųjų sąrašo (46 kintamųjų) buvo pašalinti 15 kintamųjų.

6 lentelė. Kintamųjų atrinkimo metodų analizės apibendrinimas (sudaryta autoriaus)

Šaltinis	**Visi naudoti metodai**	**Metodų taikymas ir įvertinimas**
M. Xu ir C. Zhang‘as (2009) [11], L. Cultrera ir X. Bredart‘as (2016) [13], D. Alaminos‘as, A. Castill‘as ir kt. (2016) [19], B. Asgarnezhad‘as ir M. Soltani (2016) [20]	Literatūros apžvalga	Atrenkami panašūs kintamieji
C. F. Tsai‘us (2009) [9]	T-kriterijus; Pažingsninis; Koreliacijų matrica; Faktorinė analizė; Pagrindinių komp. analizė	Priklausomai nuo duomenų rinkinių: T-kriterijus – pašalino 15%-40% kintamųjų; Faktorinė analizė – pašalino 20-30% kintamųjų
E. Fedorova ir E. V. Gilenko (2013) [5]	Logistinė regresija; Diskriminantinė analizė; Atsitiktinis miškas	Geriausias vienas metodas - logistinė regresija; Geresni rezultatai pasiekiami kombinuojant visus metodus ir pasirenkant jų atrinktus bendrus kintamuosius.
S. Tian’as, Y. Yu ir kt. (2015) [8]	Lasso metodas	-
S. Rabiul‘as, I. W. Eberle ir kt. (2019) [22]	Koreliacijų matrica; Atsitiktinis miškas; Genetinis algoritmas	Visi metodai taikomi paeiliui.

Nors dauguma autorių kintamųjų atsirinkimą grindė literatūra, keli tyrėjai pasiūlė pakankamai efektyvius būdus, taikant įvairius algoritmus. Geriausi rezultatai buvo pasiekti, kuomet kintamieji atrenkami kelių iteracijų metu – sudarant kelis modelius ar taikant kelis metodus ir nustatant vidutinę kintamųjų svarbą visų iteracijų metu. Taip pat beveik visi tyrėjų nagrinėjami kintamieji buvo pateikiami ketvirtį ar metus.

1.4. Bankroto prognozės problematika

Bankroto prognozavimo mokslinėje literatūroje yra minimos dvi kategorijos metodų [27]:

- Klasikiniai statistiniai: logistinė regresija, diskriminantinė analizė, daugianarė diskriminantinė analizė ir Z – įvertis, logit ir probit modeliai;
- Mašininio mokymo: neuroniniai tinklai, atraminės vektorių mašinos, sprendimų medžiai, genetiniai algoritmai, fuzzy, griežtų aibių.

Tačiau nepriklausomai, kurie metodai yra taikomi, taip pat reikia įvertinti ir duomenų charakteristikas – kiek kokio tipo kintamųjų yra duomenyse bei koks klasių santykis [12].

1.4.1. Klasių disbalanso problemos analizė

Sprendžiant klasifikavimo uždavinį dažnai neatsiejama yra ir klasių disbalanso problema. Dirbant su realiais duomenimis, klasės retai kada būna pasiskirsčiusios tolygiai, todėl atsiranda tikslas sukurti tokį klasifikatorių, su kuriuo pavyktų pasiekti aukštą tikslumą mažumos klasei, per daug nenukenčiant dominuojančios klasės tikslumui [2]. Literatūroje yra išskiriami keturi pagrindiniai metodai, skirti darbui su nesubalansuotais duomenų rinkiniais [1]:

1. Duomenų parinkimo metodai (angl. *sampling methods*). Šie metodai yra taikomi prieš atliekant klasifikavimą. Iš turimo duomenų rinkinio sukuriamas toks, kurio klasės būtų pasiskirsčiusios maždaug vienodai. Toks metodas nereikalauja modifikuoti klasifikavimo algoritmų ir leidžia naudoti klasikinius metodus. Yra išskiriamos pora pagrindinių duomenų perrinkimo strategijų:
    - Dirbtinė mažumos sukūrimo technika (angl. SMOTE – *synthetic minority oversampling technique*).
    - Dominuojančios klasės duomenų išmetimo (angl. *undersampling*);
    - Mažumos klasės duomenų priauginimo (angl. *oversampling*).
2. Baudomis grįsti metodai (angl. *cost-sensitive methods*). Šie metodai reikalauja tiek duomenų modifikavimo, priskiriant baudas klasėms, tiek klasifikavimo algoritmo modifikavimo, kad algoritmas atsižvelgtų į baudas. Tikslas yra suteikti didesnę baudą klaidingai suklasifikuotoms mažosios klasės reikšmėms, taip siekiant išlaikyti vienodą svarbą tarp klasių klasifikuojant.
3. Branduoliu grįsti metodai (angl. *kernel-based methods*). Šių metodų esmė yra klasifikuojant tobulinti branduolio funkciją, suteikiant svorius ir keičiant parametrus.
4. Aktyvaus mokymosi metodai (angl. *active learning methods*). Šių metodų principas yra panašus į duomenų parinkimo metodų. Algoritmas atrinkinėja taškus klasifikatorių apmokymui, taip, kad būtų išlygintas disbalansas tarp klasių ir į apmokymo imtį būtų įtraukti didžiausią informaciją nešantys taškai [3].

Tyrėjai Shaza M. Elrahman ir Ajithas Abrahamas savo tyrime palygino klasės disbalanso sprendimo metodus ir nustatė jų pranašumus ir trūkūmus. Jų apibendrintos įžvalgos pateikiamos 7 lentelėje:

7 lentelė. Įvairių metodų pranašumai ir trūkumai, sprendžiant klasių disbalanso problemą [28]

Metodas	Pranašumai	Trūkūmai
Duomenų išmetimas (undersampling)	• Nepriklausomi nuo naudojamo klasifikatoriaus; • Gali būti nesunkiai įgyvendinti	• Gali būti prarasti svarbūs duomenų bruožai, nešantys svarbią informaciją
Duomenų priauginimas (oversampling)		• Gali trukti daugiau laiko; • Persimokymo pavojus
Baudomis grįsti (cost-sensitive)	• Mažesnė žala suklasifikavus klaidingai	• Tikroji klaidingo klasifikavimo žala išlieka nežinoma
Atpažinimu grįsti	• Geresnis bendras efektyvumas, ypač su aukštos dimensijos duomenimis	• Daug klasifikatorių, tokių kaip sprendimų medžiai ir Naive Bayes negali būti sukurti remiantis viena klase
Ensemble	• Geresnis efektyvumas nei pavienių klasifikatorių • Atsparesnis triukšmui	• Gali trukti daugiau laiko; • Persimokymo pavojus

Klasių disbalanso problema yra dažnai sprendžiama medicinos analitikos srityje. Nagrinėdami retus atvejus elektroniniuose sveikatos įrašuose (angl. EHR – *electronic health records*), tyrėjai Zoja Shui-Yee, Yangas Zhao ir Kwok‘as Leung‘as Tsui‘as sukūrė sistemą retų ir panašių įvykių (angl. LASA – *look-alike sound-alike*) klasifikavimui [4]. Sprendžiant klasių disbalanso problemą tyrėjai apjungė logistinę regresiją su keliomis duomenų balansavimo strategijomis. Sistema buvo sudaryta iš dviejų dalių – pagal recall klasifikavimo metriką parenkamas bazinis binarinis klasifikatorius: logistinė regresija, atramos vektorių mašina arba sprendimų medžiai. Parinkus klasifikatorių, taikomos balansavimo strategijos: duomenų perrinkimo (angl. *resampling*) arba jautraus mokymo. Parenkant metodų parametrus buvo naudojamas kryžminis patikrinimas, vertinant klasifikatorių kokybę. Darbui atlikti buvo naudojamas R paketas su „MASS“, „e1071“, „cvTools“, „plyr“, „DmwE“ bei „tree“ bibliotekomis. Gauta, kad klasifikavimas atliekamas tiksliausiai taikant logistinę regresiją su dirbtine mažumos pavyzdžių sukūrimo technika (angl. SMOTE – *synthetic minority oversampling technique*), recall metrika siekia 75.7%, kai taikant tik logistinę regresiją tikslumas siekė 52.1%.

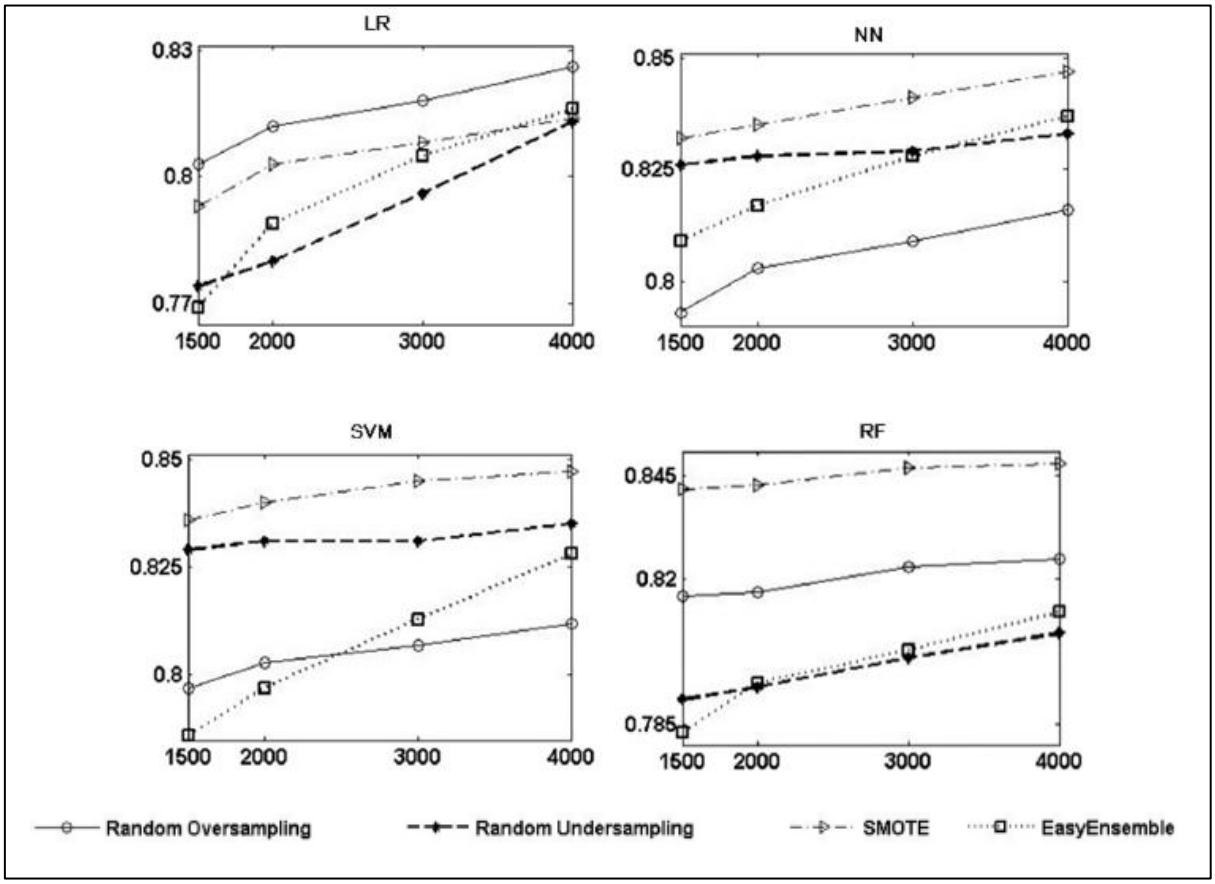
Tokiu pačiu principu – klasifikavimo ir duomenų parinkimo algoritmų sąveikos tyrimą atliko ir nagrinėjo tyrėjai Davidas Vegazonesas ir Eric‘as Severinas [7]. Jie savo darbe nagrinėjo įmonių bankroto prognozės uždavinį, kai duomenų rinkinį sudarė 1500 įmonių iš kurių 75 yra bankrutavusios. Klasių disbalanso problemai spręsti buvo naudojami atsitiktinio dauginimo, atsitiktinio išmetimo, SMOTE ir EasyEnsemble metodai. Klasifikavimo tikslumui nusakyti naudojamos G-vidurkio, AUC metrikos. Gauti rezultatai pateikti 8 lentelėje.

8 lentelė. Klasifikavimo modelių tikslumai pagal duomenų parinkimo metodus [7]

Modelis \ Metodas	Metrika	RO	RU	SMOTE	EE
LR	**AUC**	0.803	0.774	0.793	0.769
	G-vidurkis	76.4	72.7	74.3	72.6
NN	**AUC**	0.793	0.826	0.832	0.809

	G-vidurkis	73.9	76.6	79.8	75.4
SVM	**AUC**	0.797	0.829	0.836	0.786
	G-vidurkis	74.8	77.0	76.6	74.0
RF	**AUC**	0.816	0.791	0.842	0.783
	G-vidurkis	77.3	74.2	78.7	75.1

Galima pastebėti, kad SMOTE duomenų parinkimo metodas, pagal dvi metrikas yra beveik efektyviausias naudojant 3 iš 4 klasifikavimo algoritmų. Didžiausias tikslumas buvo pasiektas naudojant atsitiktinio miško ir SMOTE metodus. Taip pat tyrėjai išbandė klasifikatorių ir duomenų parinkimo metodų efektyvumą su skirtingų dydžių duomenų imtimis (žr. 3 pav.).



3 pav. Klasifikavimo algoritmų ir duomenų parinkimo metodų AUC metrikos pagal imties dydį. X ašis – imties dydis, Y ašis – AUC reikšmė [7].

Galima pastebėti, kad pagal skirtingus duomenų imties dydžius, beveik visada SMOTE algoritmas yra efektyviausias.

Mokslininkų iš Malaizijos tyrimo rezultatai [29] prieštarauja anksčiau analizuotų tyrimų rezultatams. Nagrinėdami kreditinių kortelių sukčiavimo duomenis, tyrėjai taip pat lygino SMOTE, atsitiktinio išmetimo ir atsitiktinio dauginimo metodus. Buvo nustatyta, kad taikant SMOTE tyrimo metu buvo gauti prastesni rezultatai nei taikant atsitiktinio dauginimo metodą. Tiesa, rezultatai buvo panašūs.

Kiek kitokį sprendimą pasiūlė Kinijos mokslininkai. Sprendžiant klasių disbalanso problemą buvo modifikuotas atraminės vektorių mašinos algoritmas, nepertvarkant pradinių duomenų [6]. Algoritmo modifikacija grįsta tinkamos hiperplokštumos parinkimu, kuris yra suvestas į iteracinį procesą. Pirmiausia taikomas klasikinis SVM metodas siekiant parinkti pradinę hiperplokštumą, vėliau taikant chi-kvadrato kriterijų parenkami parametrai ir svoriai (duomenys, priklausantys dominuojančiai klasei gauna mažesnį svorį nei priklausantys mažajai klasei). Taip sukonstruojama nauja hiperplokštumos funkcija ir vykdomas klasifikavimas. Ši algoritmo modifikacija buvo sukurta spręsti tiek dviejų klasių, tiek daugiau klasių klasifikavimo problemą. Išbandžius algoritmą su trejais duomenų rinkiniais bei lyginant su kitais algoritmais: klasikiniu SVM, bendru regresiniu neuroniniu tinklu (angl. GRNN – *general regression neural network),* tikimybiniu neuroniniu tinklu (angl. PNN – *probabilistic neural network*). Bandymų rezultatai, tikslumo metrikos, pateiktos 9 lentelėje.

9 lentelė. Algoritmų tikslumas su skirtingais duomenų rinkiniais [29]

Duom. rinkinys	**klasikinis SVM**	**GRNN**	**PNN**	**Modifikuotas SVM**
Haberman	65.63	68.75	67.71	86.46
Glass	82.61	75.36	82.61	86.96
Ecoli	89.47	86.84	88.60	92.98

Gaunama, kad visais atvejais pasiūlytas modifikuotas SVM algoritmas pasiekia 86.46% – 92.98% tikslumą, kas yra geriausias rezultatas lyginant su kitais algoritmais. Galima pastebėti, kad duomenų rinkiniai, su kuriais buvo bandomas algoritmas, neturėjo labai didelio disbalanso (disbalanso santykis siekė tik ~3 – 8), tačiau tyrėjai parodė, kad galima efektyviai spręsti tokias klasifikavimo problemas nepertvarkant duomenų, o modifikuojant klasifikavimo algoritmo veikimą.

Panašiai – nepertvarkant duomenų, o pertvarkant egzistuojančių algoritmų veikimą - klasių disbalanso problemą sprendė Yusuf Sahin [30] bei Masoumeh Zareaapoor [31]. Abiejuose tyrimuose buvo naudojami kreditinių kortelių duomenys, siekiant nustatyti sukčius. Abu autoriai pabrėžė, kad atsitiktinio miško algoritmas klasifikuojant yra labiau atsparus klasių disbalansui nei kiti algoritmai.

Paprastesnį metodą klasių disbalansui spręsti taip pat panaudojo mokslininkai iš Rusijos [5]. Jų nagrinėjamą duomenų imtį sudarė 3056 įmonės iš kurių bankrutavusios buvo 444. Sprendžiant klasių disbalanso problemą, tyrėjai panaudojo duomenų perrinkimo metodą ir iš visos imties darbui atsitiktinai parinko 444 nebankrutavusias įmones.

10 lentelė. Klasių disbalanso problemos sprendimo literatūros apžvalgos apibendrinimas (sudaryta autoriaus)

Šaltinis	**Naudoti metodai**	**Geriausias metodas ar jų sąveika**
Z. S. Yee, Y. Zhao ir kt. (2018) [4]	Duomenų perrinkimas Jautrus mokymas	SMOTE + logistinė regresija
D. Vegazones‘as ir E. Severin‘as (2018) [7]	Atsitiktinio dauginimo Atsitiktinio išmetimo SMOTE EasyEnsemble	SMOTE + atsitiktinis miškas

H. Nur‘as, S. Siti‘ja ir kt (2018) [29]	Atsitiktinio dauginimo Atsitiktinio išmetimo SMOTE	Atsitiktinis dauginimas, panašus tikslumas buvo pasiektas taikant ir SMOTE metodą
Y. Zhang‘as, P. Fu‘as ir kt. (2014) [6]	Modifikuotas SVM algoritmas	-
S. Yusuf‘as, B. Serol‘as ir kt. (2013) [30], Z. Masoumeh‘as ir S. Pourya (2015) [31]	Atsitiktinio miško Atsitiktinio išmetimo	Klasių disbalanso problema klasifikuojant buvo sprendžiama keičiant disbalansą apmokymo imtyje. Atsitiktinis išmetimas
E. Fedorova ir E. V. Gilenko (2013) [5]	Atsitiktinio išmetimo	-

Atlikus klasių disbalanso literatūros analizę, galima pastebėti, kad yra nemaža disbalanso sprendimo būdų įvairovė. Vienas efektyviausių naudojamų algoritmų yra SMOTE, taip pat labai plačiai naudojamas paprastesnis ir spartesnis – atsitiktinio išmetimo metodas.

1.4.2. Klasifikavimo metodų taikymo bankroto prognozei tyrimų analizė

Toliau nagrinėjant Rusijos įmonių bankroto prognozės problemą [5], klasifikavimui buvo taikomi daugiasluoksniai ir radialine funkcija grįsti neuroniniai tinklai. Gautos tikslumo metrikos kiekvienai kintamųjų grupei ir kintamųjų skaičiai pateikti 11 lentelėje.

11 lentelė. Klasifikavimo tikslumo palyginimas naudojant skirtingus kintamuosius ir neuroninių tinklų algoritmus [5]

Modelis	**Kint. Sk.**	**Bendras tikslumas**	**Precision**	**Sensitivity**	**Specificity**	**F-measure**
Daugiasluoksnis neuroninis tinklas						
MDA	13	74.5	80.0	65.3	83.7	71.9
CRT	**6**	**86.7**	**86.0**	**87.8**	**85.7**	**86.9**
LR	**8**	**87.8**	**86.3**	**89.8**	**85.7**	**88.0**
118-MinEcon	3	70.4	71.7	67.3	73.5	69.5
367-GovRF	5	84.7	80,4	91.8	77.6	85.7
Radialine funkcija grįstas neuroninis tinklas						
MDA	13	70.4	73.8	63.3	77.6	68.1
CRT	6	78.6	75.0	85.7	71.4	80.0
LR	8	80.6	82.6	77.6	83.7	80.0
118-MinEcon	3	70.4	71.7	67.3	73.5	69.5
367-GovRF	5	78.6	78.0	79.6	77.6	78.8

Šiuo atveju geriausias tikslumas buvo pasiektas taikant logistinę regresiją kintamųjų atrinkimui ir daugiasluoksnį neuroninį tinklą klasifikavimui, tikslumas – 87.8%. Toliau patobulinus kintamųjų atrinkimo procesą, geriausias tikslumas buvo pasiektas taip pat taikant daugiasluoksnį neuroninį tinklą – tikslumo metrika siekė 88.8%. Straipsnio autoriai pabrėžė reikšmingų kintamųjų atrinkimo svarbą, siekiant prognozuoti įmonės bankrotą.

Neuroninius tinklus įmonių bankroto prognozavimui taip pat naudojo jau anksčiau apžvelgto straipsnio mokslininkai [9]. Kintamųjų atrinkimui panaudojus t-kriterijų, bendras klasifikavimo tikslumas siekia 70% - 90%. Tyrimo apibendrinimas pateiktas 12 lentelėje.

12 lentelė. Klasifikavimo rezultatai, kintamųjų atrinkimui taikant T-kriterijų [9].

Duomenys	**Disbalansas**	**Duomenų kiekis**	**Prad. kint. kiekis**	**Tikslumas**	**I tipo klaida (veik. -> bankr.)**	**II tipo klaida (bankr. -> veik.)**
Japonijos	0.55	690	15	85.88 %	90.05 %	22.4 %
Australijos	0.55	690	14	81.93 %	21.89 %	13.89 %
Bankroto	0.47	240	33	71.03 %	12.85 %	30.42 %
Vokietijos	0.3	1000	20	74.28 %	55.39 %	9.63 %
UC	0.03	2528	39	96.92 %	81.68 %	4.05 %

Galima pastebėti, kad nors bendras tikslumas atrodo neblogas, I ir II tipo klaidų yra nemažai. Algoritmas yra dažniau linkęs veikiančią įmonę priskirti prie bankrutavusių įmonių. Taip pat reiktų atkreipti dėmesį, kad duomenų imtys ir disbalansas nėra dideli.

Su kiek didesne imtimi tyrimą atliko Belgijos mokslininkai [13]. Jų duomenų imtį sudarė 7152 įmonės – pusė jų buvo bankrutavusios. Bankroto prognozavimui buvo taikoma logistinė regresija. Sukurtas logistinės regresijos modelis klasifikavo įmones 79.23% tikslumu.

Logistinę regresiją taip pat taikė mokslininkai, tirdami įmones iš Kipro akcijų biržos [20]. Čia bankrutavusių įmonių buvo 0.23 visos imties, o bendras klasifikavimo tikslumas siekė net 91.8%.

Be mašininio mokymo algoritmų, mokslininkai taip pat bandė taikyti klasikinius statistinius metodus. Anksčiau apžvelgtame straipsnyje [11] tyrėjai lygino Altmano Z-įverčio, Ohlsono O-įverčio ir apjungtus metodus įmonių bankroto klasifikavimui. Darbe buvo naudojami duomenys nuo 1992 iki 2005 metų. Duomenų imtį iš viso sudarė 3510 Japonijos įmonių iš įvairių sektorių, iš kurių bankrutavusios buvo 76. Naudoti modeliai klasifikavo 59.2% - 72.4% tikslumu.

Tyrėjas iš Italijos Massimilianas Celi taip pat taikė Altmano Z-įverčio metodą siekiant suklasifikuoti bankrutavusias įmones [15]. Bandymo duomenų imtį sudarė 102 Italijos įmonės iš kurių 51 buvo bankrutavusi laikotarpyje nuo 1995 iki 2013 metų. Modelis buvo atskirai taikomas bankrutavusioms kompanijoms ir gyvuojančioms. Taip pat buvo bandomi skirtingi laikotarpiai iki bankroto: 1, 2 ir 3 metai. Tiriant bankrutavusias įmones, buvo gauta, kad metams iki bankroto teisingai suklasifikuotos 84.3% įmonių, dvejiems metams - 70.5%, trims metams - 47.1% įmonių. Nagrinėjant gyvuojančias įmones klasifikatorius buvo tikslesnis, atitinkamai klasifikavimo tikslumai buvo 90.1%, 84.3% ir 86.2%. Modelio tikslumas buvo didžiausias naudojant paskutinių metų įmonių finansinius duomenis.

13 lentelė. Klasifikavimo metodų taikymo mokslinės literatūros analizės apibendrinimas (sudaryta autoriaus)

Šaltinio nr.	Disbalanso koef. (bakrutavusios/visos)	Geriausias / taikytas klasifikavimo metodas	Ar spręsta disbalanso problema?	Tikslumo metrikos
Z. S. Yee, Y. Zhao‘as ir kt. (2018) [4]	0.211	Logistinė regresija	SMOTE	Recall, ROC, bendras tikslumas, F-matas, I ir II tipo klaidos
E. Fedorova ir E. V. Gilenko (2013) [5]	0.144	Neuroniniai tinklai	Atsitiktinis išmetimas	Bendras tikslumas, specifiškumas, jautrumas.
D. Vegazones‘as ir E. Severin‘as (2018) [7]	0.05	Atsitiktinis miškas	SMOTE	ROC, G-vidurkis, bendras tikslumas
C. F. Tsai‘us (2009) [9]	0.03 – 0.55	Neuroniniai tinklai	Ne	Bendras tikslumas, I ir II tipo klaidos
M. Xu ir C. Zhang‘as (2009) [11]	0.022	Altmano Z- įvertis Ohlsono O-įvertis	Ne	Bendras tikslumas
L. Cultrera ir X. Bredart‘as (2016) [13]	0.5	Logistinė regresija	Ne	Bendras tikslumas, I ir II tipo klaidos
M. Celli (2015) [15]	0.5	Altmano Z-įvertis	Ne	Bendras tikslumas
S. Yusuf‘as, B. Serol‘as ir kt. (2013) [30]	0.1	Atsitiktinis miškas	Atsitiktinis išmetimas	Teisingi-teigiami atvejai
Z. Masoumeh‘as ir S. Pourya (2015) [31]	0.03 – 0.2	Atsitiktinis miškas	Atsitiktinis išmetimas	Balansuotas tikslumas, Klaidingi-teigiami atvejai

Galima pastebėti, kad dirbant su didesniu klasių disbalansu, plačiausiai taikomi yra atsitiktinių miškų, neuroninių tinklų ir logistinės regresijos algoritmai. Dėl parametrų parinkimo ir derinimo galimybių, veikimo spartos ir rezultatų interpretavimo daugiausiai buvo naudojami atsitiktinių miškų algoritmai. Taip pat klasifikavimo algoritmai apjungiami su disbalanso sprendimo metodais, siekiant kokybiškiau apmokyti modelius.

1.4.3. Klasifikavimo modelių vertinimas ir tikslumo metrikos

Sudarant klasifikavimo modelius duomenims su dideliu disbalansu yra svarbu parinkti tinkamą klasifikavimo vertinimo metriką [32]. Naudojant įprastas metrikas atsiranda pavojus klaidingai interpretuoti rezultatus ir pagal tai priimti blogus sprendimus, kadangi dažnai dalis metrikų neatsižvelgia į klasių proporcijas ir yra linkusios vertinimą pakreipti pagal dominuojančią klasę.

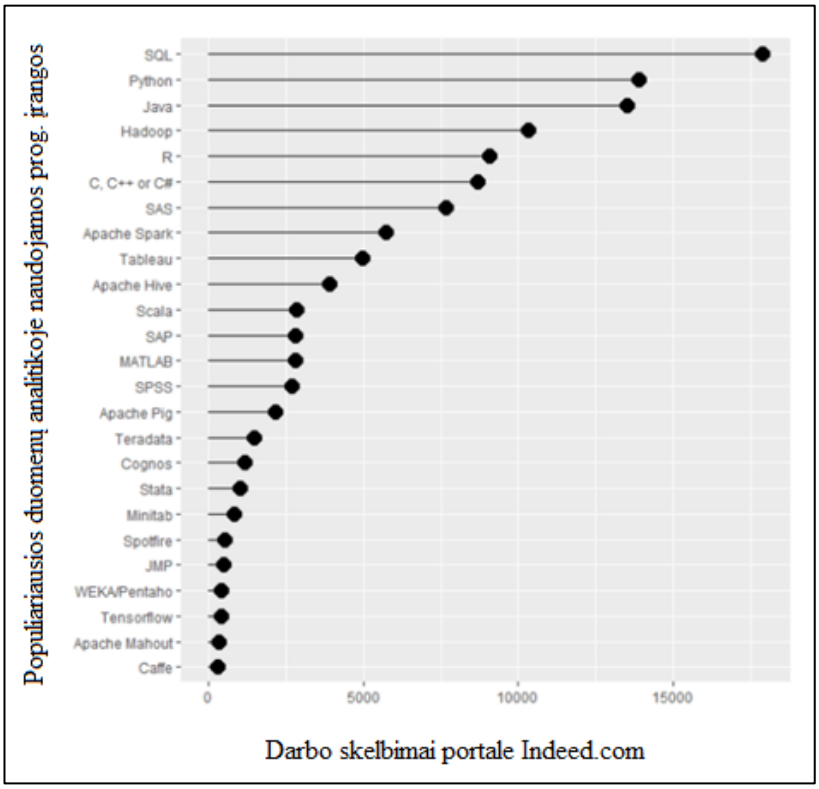
Tyrėjai Janas Brabecas ir Lukas Machilas [33], besiremdami kitų mokslininkų darbais atliko klasifikavimo metrikų apžvalgą ir nustatė, kad nemaža dalis mokslininkų savo darbuose naudojo klaidingas metrikas, nors duomenyse buvo didelis disbalansas. Tyrėjai rekomendavo naudoti tikslumo (angl. *precision*), atpažinimo (angl. *recall*) bei F1-įverčio metrikas. Taip pat siūlomi būdai modelį vertinti grafiškai: pagal gerai ir blogai suklasifikuotų teigiamos klasės atvejų santykį, kurį nusako ROC kreivė (angl. ROC - *Receiver Operating Characteristic*), tikslumo – atpažinimo (angl. *Precision - Recall*) kreivė bei DET (angl. DET - *Detection Error Tradeoff*) kreivė.

Kitas tyrėjas – Alaa Tharwata [34] plačiau aprašo dažniausiai naudojamas metrikas ir taip pat pabrėžia klasifikavimo metrikų pasirinkimo svarbą, atsižvelgiant į klasių disbalansą duomenyse. Klasifikavimo vertinimui, kai darbui buvo naudojami duomenys su disbalansu, tyrėjas išskiria tas pačias metrikas kaip ir anksčiau apžvelgtame straipsnyje. Jis taip pat prideda, kad geometrinis vidurkis (angl. *geometric mean*) arba koreguotas geometrinis vidurkis (angl. *adjusted geometric mean*) taip pat gali būti naudojamas klasifikavimo vertinimui.

1.5. Programinės įrangos apžvalga

Apžvelgtuose straipsniuose beveik nebuvo minimos programinės įrangos, kurios buvo naudojamos atliktuose tyrimuose, tačiau mokslininkai M. F. Delagas ir E. Cernadas palygino 179 skirtingus klasifikatorius taikydami 121 duomenų rinkinį [37]. Naudoti klasifikatoriai buvo iš įvairios programinės įrangos: *Weka, R, C, Matlab*. Geriausias vidutinis tikslumas buvo pasiektas taikant *R* pakete esančius atsitiktinių miškų algoritmus, antras pagal gerumą buvo *C* programavimo kalba parašytas atraminių vektorių mašinų klasifikatorius.

Paskutiniu metu taip pat sparčiai populiarėja *Python* programavimo kalba. Šią, *R* ir *Matlab* palygino tyrėjai C. Ozguras ir kiti [38]. Tyrėjai pabrėžia, kad *Matlab* pirmiausia buvo kurta kaip įrankis dirbti ir manipuliuoti matricomis ir tik vėliau buvo ištobulinta, sukuriant savo programavimo kalbą. Vizualizacijų kūrimas bei patogus darbas su integralais ir matricomis yra išskiriami kaip didžiausi *Matlab* pranašumai. Tačiau didžiųjų duomenų analitikoje tyrėjai rekomenduoja naudoti *R* ir *Python* programavimo kalbas, pagrįsdami tuo, kad abi kalbos yra nemokamos, joms yra sukurta daug papildymų bei bibliotekų ir jos yra populiariausios darbo rinkoje (žr. 4 pav.).



4 pav. Darbo skelbimai pagal populiariausią duomenų analitikoje naudojamą programinę įrangą

Galiausiai tyrėjų rekomendacija yra pirmiau išmokti *R*, kadangi ji paprasčiau įvaldyti kodo rašymo atžvilgiu, ji lengvai taikoma mažesniems projektams. *Python* kalba, nors ir populiaresnė, šalia statistikos ir matematikos reikalauja daugiau programavimo žinių.

Nagrinėjant *R* paketo taikymą, sprendžiant klasifikavimo problemą, tyrėjai M. N. Wright'as ir A. Ziegler'is sukūrė *ranger* (angl. RANGER - *RANdom forest GEneRator*) paketą, skirtą atsitiktinių miškų taikymui didelių dimensijų duomenims [39]. Tyrėjai pritaikė jų anksčiau sukurtą *ranger* paketą iš *C++* aplinkos, jį praplėtė ir sukūrė išsamią dokumentaciją. Su tuo pačiu duomenų rinkiniu *ranger* paketą palyginę su kitomis *R* įgyvendintomis atsitiktinio miško alternatyvomis, buvo nustatyta, kad *ranger* paketas yra efektyviausias, kai dirbama su daug didelių dimensijų duomenimis. Po metų nuo šio paketo sukūrimo, tas pats mokslininkas su kitų tyrėjų pagalba praplėtė *ranger* taikymą *R* pakete ir sukūrė automatinio hyper-parametrų derinimo paketą *tuneRanger*, kuris atsitiktinių miškų taikymą padarė dar efektyvesnį [40].

1.6. Tyrimo uždaviniai ir pasirinktų metodų pagrindimas

Bendradarbiaujant su valstybine mokesčių inspekcija bei Sodra, buvo iškeltas tikslas ištirti kokią įtaką duomenų specifika turi bankroto procedūros prognozei ir nustatyti, kaip tikslingiausia panaudoti turimus duomenis, norint kokybiškai prognozuoti ar įmonei bus pradėta bankroto procedūra.

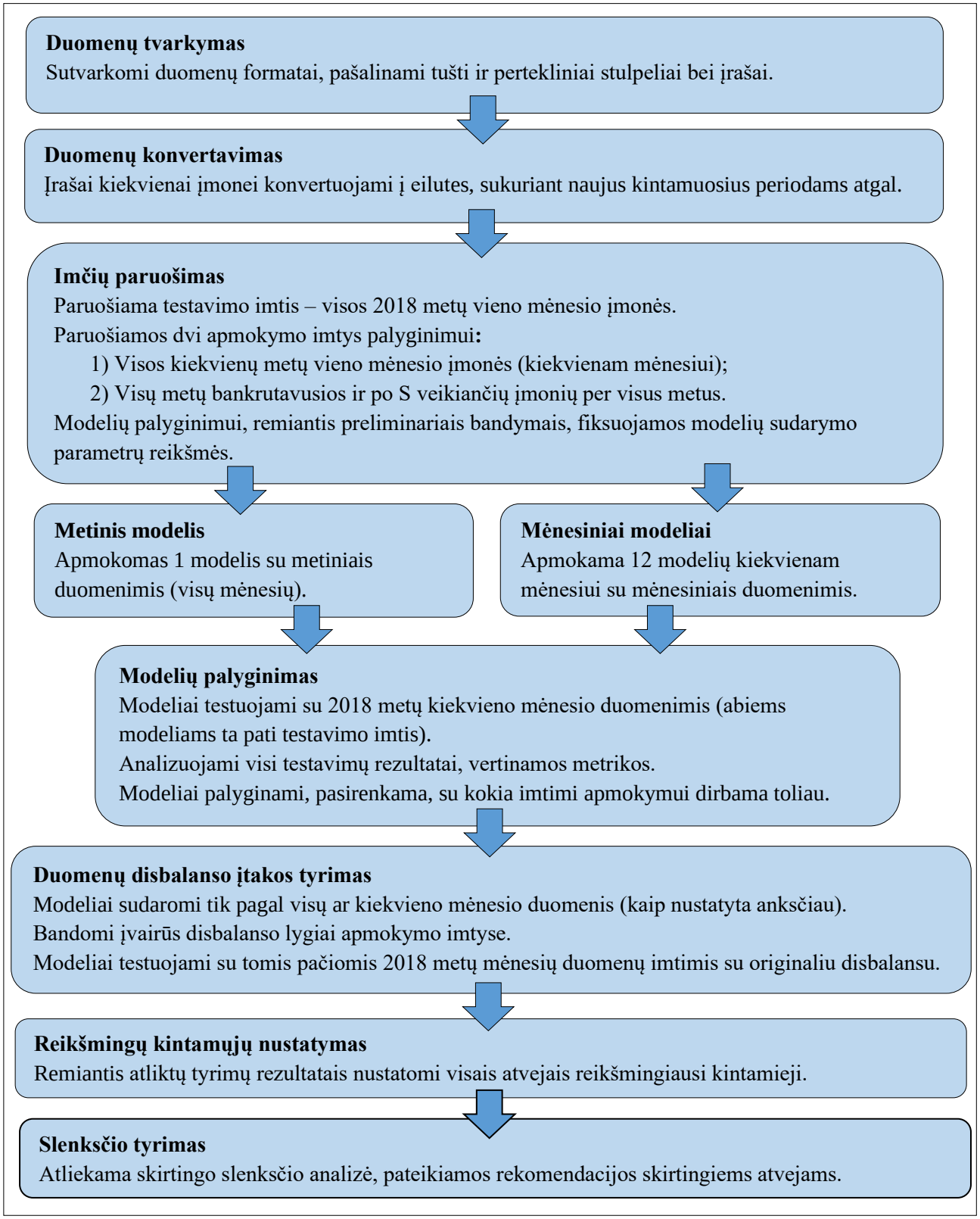
Atlikus literatūros analizę, galima pastebėti, kad bankroto prognozės uždaviniui spręsti yra plačiai naudojami mašininio mokymo algoritmai. Taip pat, dirbant su finansiniais įmonių duomenimis bankroto procedūros prognozei, dažnai yra susiduriama su disbalanso problema. Literatūroje dažniausiai nagrinėjami duomenys yra ketvirtiniai arba metiniai, taip pat analizuotose straipsniuose nevertinama rodiklių, susijusių su darbuotojais įmonėje, įtaka bankroto procedūrai.

Atsižvelgiant į literatūros apžvalgą, dėl gero rezultatų ir veikimo spartos santykio bei parametrų įvairovės ir jų derinimo metodų pasirinktas mašininio mokymo atsitiktinių miškų metodas. Taip pat taikant atsitiktinių miškų metodą bei statistinę analizę bus atrinkti reikšmingiausi kintamieji. Disbalansui kontroliuoti tyrime bus naudojamas atsitiktinio išmetimo algoritmas.

Kadangi *R* paketas yra nemokamas, turi didelį bibliotekų pasirinkimą ir yra viena iš rekomenduojamų programinių įrangų sprendžiant klasifikavimo problemą bei dirbant su didelės apimties duomenimis, jis bus naudojamas tyrime. Atsitiktinių miškų taikymui pasirenkamas neseniai sukurtas, sparta pasižymintis *ranger* paketas.

2. Tyrimo metodai

Skyriuje pateikti tyrime naudojami metodai. Kadangi tyrimą apima keli etapai, sudaromas tyrimo planas (žr. 5 pav.).

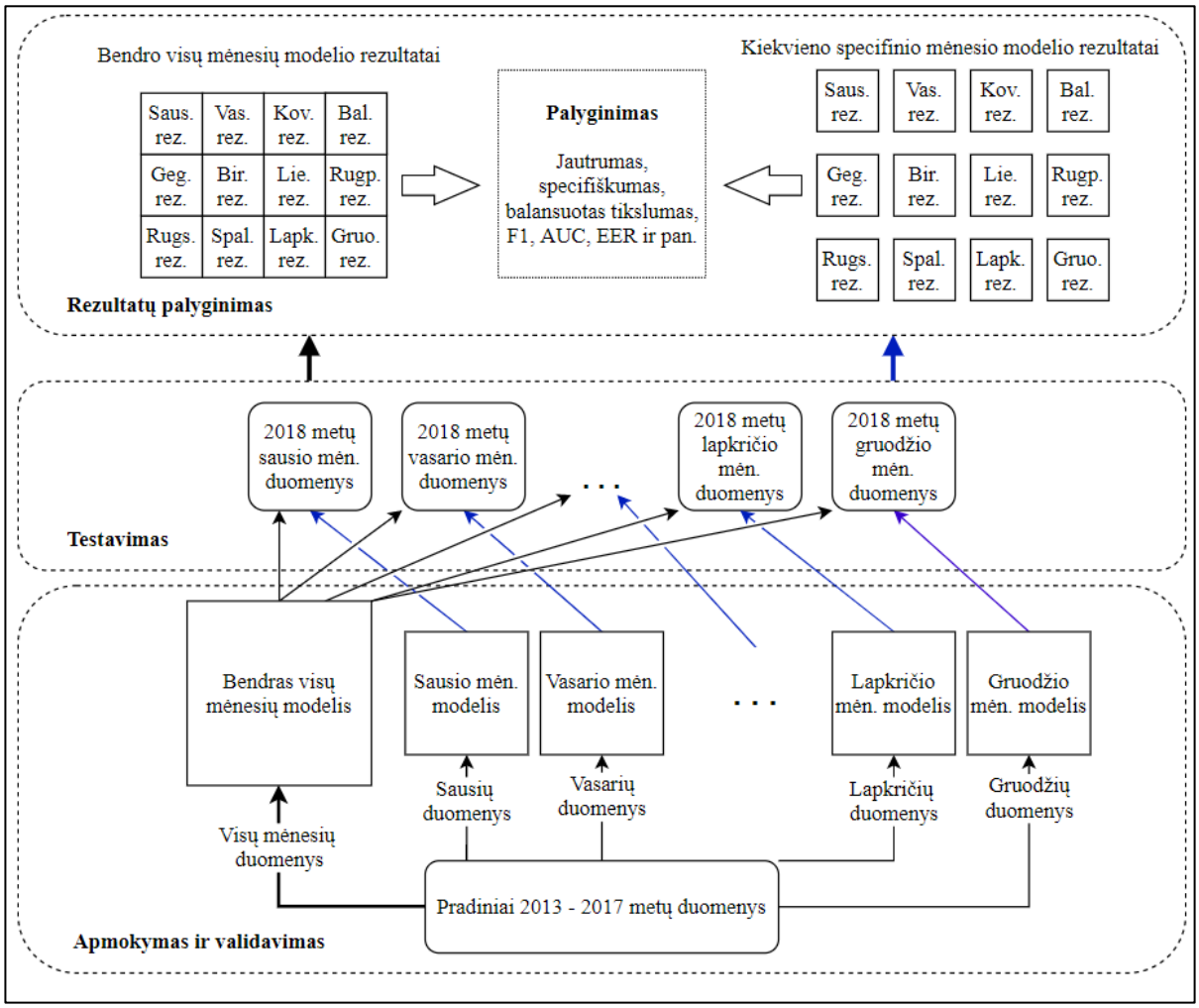


5 pav. Tyrimo eigos schema

Norint tinkamai pritaikyti turimus duomenis, reikia juos sutvarkyti ir konvertuoti. Pirmame tyrimo etape stulpeliai ir eilutės yra išvalomi, pagal įvairius kriterijus (kurie detaliau apžvelgiami toliau), duomenys konvertuojami į kitą formatą.

Kadangi vienas iš uždavinių yra palyginti vieną bendrą visų mėnesių modelį su kiekvieno mėnesio specifiniais modeliais – sukuriama 13 modelių apmokymo duomenų imčių. Bendram visų mėnesių modeliui sudaroma 1 imtis, į kurią patenka visos įmonės, kurioms per visų 2013 – 2017 metų mėnesius buvo pradėta bankroto procedūra ir kiekvienam bankroto atvejui atitinkamas skaičius veikiančių įmonių. Kiekvienam specifiniam mėnesio modeliui sudaroma po 1 imtį, į kurią patenka visos įmonės, kurioms bankroto procedūra buvo pradėta tik tą mėnesį 2013 – 2017 metais ir kiekvienam bankroto atvejui atitinkamas kiekis veikiančių įmonių – tokiu principu sudaroma 12 imčių kiekvieno mėnesio modeliui. Pasiruošimo rezultatas – 13 duomenų imčių modeliams apmokyti su vienodu disbalanso lygiu.

Sukūrus atsitiktinių miškų modelius, modeliai testuojami su 2018 metų kiekvieno mėnesio duomenimis - gaunami 24 rezultatai. Kiekvieno mėnesio testavimo imtį sudaro visos tą mėnesį veikusios įmonės ir visos įmonės, kurioms buvo pradėta bankroto procedūra. Gauti bendro modelio 12 mėnesių klasifikavimo rezultatai palyginami su kiekvieno iš 12 specifinių mėnesinių modelių klasifikavimo rezultatais. Vertinamos jautrumo, tikslumo, balansuoto tikslumo ir kitos metrikos. Detalesnė šių etapų schema pateikta 6 pav..



6 pav. Apmokymo imčių paruošimo, modelių sukūrimo ir rezultatų palyginimo schema

Darbas tęsiamas su nustatytais geriausiais modeliais. Tiriant disbalanso įtaką apmokymo imtyje (veikiančių įmonių skaičiaus santykis su bankrutuojančių skaičiumi), sukuriama N_D apmokymo imčių, kiekvienoje jų - veikiančių įmonių skaičius vienai bankrutuojančiai yra parenkamas skirtingas. Su kiekviena imtimi modeliai yra apmokomi ir testuojami su tomis pačiomis, ankstesniame etape naudotomis mėnesinėmis testavimo imtimis.

Tiriant reikšmingiausius kintamuosius, atsižvelgiama į visų sukurtų atsitiktinių miškų modelių rezultatus – kintamųjų reikšmingumus. Įvertinamas kiekvieno kintamojo svarbos vidurkis iš visų modelių bei jų didžiausios svarbos.

Siekiant nustatyti slenksčio įtaką rezultatams ir jų interpretacijai, tos pačios testavimo imties rezultatai įvertinami, parenkant kelias skirtingas slenksčių reikšmes. Priklausomai nuo tikslo ir galimos modelio paskirties, pateikiamos slenksčio rekomendacijos. Taip pat atsižvelgiama į duomenų disbalansą apmokymo imtyje.

2.1. Tyrime naudojami duomenys

Pradinę duomenų imtį sudaro duomenys nuo 2013-01-01 iki 2018-12-01 (imtinai) visų Lietuvos įmonių (apie 200 tūkst. įmonių) metinės ir mėnesinės finansinės ataskaitos ir deklaracijos iš valstybinės mokesčių inspekcijos ir Sodros. Duomenys surinkti iš PVM deklaraciją, pelno - nuostolių ataskaitų, metinių deklaracijų. Visos įmonės yra nuasmenintos, kiekvieną įmonę atitinka unikalus 13 skaitmenų kodas.

Visą duomenų rinkinį sudaro 6 failai. Kiekvieną duomenų failą sudaro nuo 1.101.960 iki 1.199.484 eilučių ir 58 kintamieji. Bendras visų duomenų failų dydis 5,17 GB. Duomenys pateikti įmonių valstybinei mokesčių inspekcijai ir Sodrai. Pagrindinės kintamųjų grupės:

- Bendra informacija apie įmonę:
  - įmonės amžius;
  - ekonominės veiklos rūšis – 3 kint.;
  - veiklos lokacija – 2 kint.;
  - mokesčių mokėtojo informacija – 3 kint.;
  - teisinis statusas ir kodas – 3 kint.
- Kintamieji iš metinių deklaracijų:
  - rodikliai apie pardavimus – 5 kint.;
  - įmonių turto ir įsipareigojimo rodikliai – 7 kint.;
  - išvesti santykiniai kintamieji – 10 kint.;
- Kintamieji iš mėnesinių deklaracijų:
  - rodikliai apie darbuotojus – 11 kint.;
  - pvm duomenys – 6 kint.;
  - mokestinių nepriemokų informacija – 2 kint.

2.1.1. Duomenų tvarkymas

Kadangi duomenų kiekis yra didelis, norint gauti kuo tikslesnius rezultatus bei taupant resursus, reikalingas bent minimalus duomenų tvarkymas. Duomenų tvarkymas pradedamas iš visos duomenų imties pašalinant įrašus atsižvelgiant į dalykinės srities ekspertų įžvalgas. Pirminio duomenų rinkinio tvarkymo etapai:

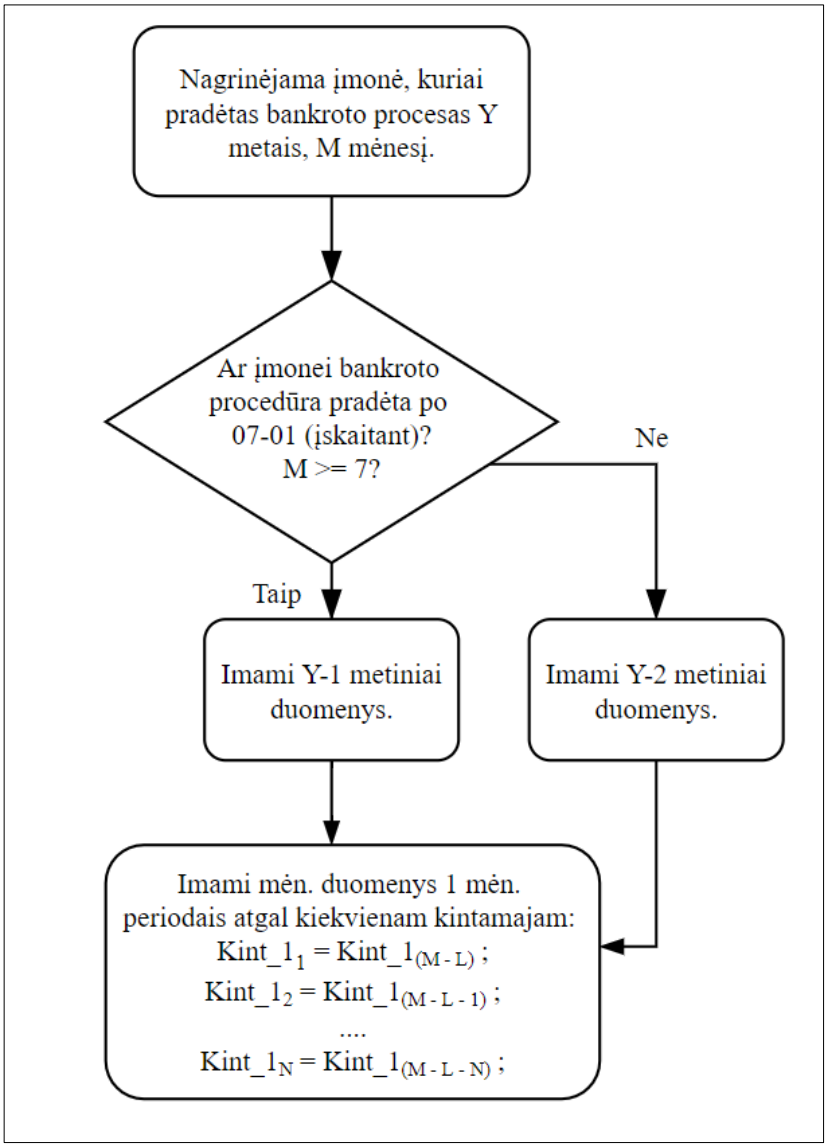
1. Pirminiai duomenys turi perteklinių stulpelių – kai kurie stulpeliai dubliuoja kitus, atkartodami identišką informaciją, tik kitokiu formatu arba indikuoja, kad kitame stulpelyje reikšmė yra nelygi 0. Tokių stulpelių paliekama po vieną - kurie duoda daugiau informacijos.
2. Duomenyse yra nemažai trūkstamų reikšmių. Pašalinami stulpeliai, kurie yra 99% tušti.
3. Kai kurie stulpeliai suformatuoti skirtingai, lyginant su kitais duomenimis. Visi 2013-2014 metų duomenys įvesti litais, todėl reikalingas konvertavimas į eurus. Taip pat kai kuriems stulpeliams naudojamas kitas dešimtainis skyriklis.
4. Pašalinami įrašai įmonių, kurios yra priskirtos „neaiškiųjų“ sekcijai – X. Apie tokių įmonių veiklą yra žinoma labai mažai, gali būti, kad jos nevykdo veiklos jau kelis metus ir jų veikla neduoda jokios naudingos informacijos apie galimą bankroto procedūros pradžią.
5. Pašalinami įrašai apie įmones, kurios susikūrė tais metais – jų gyvavimo laikotarpis yra mažesnis nei vieneri metai. Išimtis taikoma įmonėms, kurios sekančiais ar dar kitais metais bankrutavo.
6. Pašalinami įrašai apie įmones, kurios nagrinėjamais metais nepateikė metinių ataskaitų. Taip pat taikoma išimtis nagrinėjamais metais ar vėliau bankrutavusioms įmonėms.
7. Nuo pirmo bankroto procedūros iniciavimo pašalinami visi kiti vėlesni tos įmonės įrašai. Taip paskutinis likęs mėnesinis įmonės įrašas yra pirmas bankroto įrašas šiai įmonei.

2.1.2. Duomenų konvertavimas ir naudojimas bankroto prognozei

Norint prognozuoti ar įmonei bus iškelta bankroto pradžios procedūra ateinantį mėnesį, tariame, kad anksčiausia žinoma informacija yra pateikta tik užpraeitą mėnesį, kadangi tiek už einamąjį, tiek už praeitą mėnesius dar ne visos įmonės gali būti pateikusios savo finansines ataskaitas. Jeigu, pavyzdžiui, įmonei bankroto procedūra pradėta gegužės mėnesį (arba norima prognozuoti gegužės mėnesį), tuomet anksčiausias galimas mėnesinis įrašas yra tų pačių metų vasario mėnesio.

Bankrotui prognozuoti šalia mėnesinių duomenų taip pat yra naudojami metiniai. Metinius duomenis įmonės deklaruoja kiekvienų metų birželio mėnesį. Yra tariama, kad jeigu bankroto procesą norima prognozuoti liepą ir vėliau - prognozei galima naudoti praeitų metų metinius duomenis. Jeigu bankroto procesas pradėtas nuo sausio iki birželio imtinai – prognozei naudojami užpraeitų metų metiniai duomenys.

Kadangi darbe bus naudojami mašininio mokymo algoritmai, duomenis reikia apjungti ir konvertuoti į kitą, algoritmams tinkamą formatą. Gautuose duomenyse vienų metų rinkinyje viena įmonė turi 12 įrašų (vienas įrašas vienam mėnesiui). Turimus duomenis reikia suvesti į vieną įrašą vienai įmonei vienam laikotarpiui, sukuriant papildomus kintamuosius kiekvienam mėnesiui atgal. Remiantis apžvelgta literatūra ir dalykinės srities ekspertais, duomenys konvertuojami kuriant kintamuosius periodams atgal, skaičiuojant nuo datos, kada įmonei buvo pradėtas bankroto procesas, arba nuo datos, kuriai norima prognozuoti bankroto procesą. Konvertuojant duomenis, kiekvienam naudojamam kintamajam paimama N jo praeities reikšmių (čia N – praeities horizontas, kiek mėnesių imama atgal), taip sukuriant dydžius: $kint\_A_N$, $kint\_A_{N-1}$, ..., $kint\_A_1$ (žr. **7 pav.**).



7 pav. Duomenų konvertavimo schema. Čia Y, M – metai ir mėnesis kuomet įmonei buvo pradėta bankroto procedūra; Kint_1 – nagrinėjamas kintamasis; L – laiko periodas, už kiek mėnesių nuo bankroto atgal yra galimas pirmas mėnesinis įrašas; N – kiek mėnesių atgal imama.

Siekiant sudaryti kokybišką apmokymo ir validavimo duomenų imtį, renkant veikiančias įmones, yra atsižvelgiama į įmones, kurioms buvo pradėta bankroto procedūra. Nagrinėjant veikiančias įmones, turimi visų metų duomenys nuo sausio iki gruodžio – imant šių įmonių paskutinius deklaruotus duomenis, atsirastų rizika neįtraukti į modelį informacijos apie ekonominius ciklus ir įvykius rinkose – modelis gautųsi iškreiptas. Siekiant išlaikyti šią informaciją, pirmiausiai apdorojamos įmonės, kurioms buvo iškelta bankroto procedūra. Vėliau kiekvienai įmonei, kuriai buvo pradėta bankroto procedūra, pagal datą imamas tam tikras skaičius veikiančių įmonių. Taip gaunama, kad galutinėje imtyje bus proporcingas skaičius veikiančių ir su pradėta bankroto procedūra įmonių, priklausomai nuo datos, o įrašai apie veikiančias įmones kartu teiks informaciją apie tuometinį ekonominį foną – bus atsižvelgta į laikotarpį.

Kadangi darbe siekiama prognozuoti ar įmonėms bus pradėtas bankroto procesas 2018 metais, modelių apmokymo ir parametrų validavimo žingsniui bus naudojami 2013 - 2017 metų duomenys imtinai. Taip pat anksčiausias bankroto įrašas turi būti ne ankstesnis nei 2014 metų liepos mėnesį, kadangi kitu atveju jau reikėtų 2012 metų duomenų.

2.1.3. Duomenų disbalanso kontrolė

Ruošiant modelį klasifikavimui yra svarbu jį apmokyti taip, kad jis gerai klasifikuotų mažumos klasę. Šiam tikslui gali būti koreguojama apmokymo duomenų imtis. Kadangi duomenys turi didelį disbalansą, galima arba mažinti dominuojančios klasės įrašus taip prarandant informaciją, arba dirbtinai didinti mažumos klasę. Šiame tyrime yra naudojama dirbtinė mažumos kūrimo technika (angl. *SMOTE – synthetic minority over-sampling technique*).

SMOTE technikos pranašumas yra, kad naudojant ją, nauji įrašai yra ne dubliuojami, bet sukuriami, atsižvelgiant į kitus, atsitiktinai parinktus, tos klasės įrašus – taip išvengiant modelio persimokymo ir informacijos praradimo. Algoritmo veikimas yra pagrįstas tiesine interpoliacija:

1. Iš mažumos klasės stebėjimų sudaromas poaibis *A*. Kiekvienam $x \in A$ yra parenkami *K* artimiausių kaimynų apskaičiuojant Euklido atstumą tarp stebėjimo *x* ir kitų aibės *A* elementų;
2. Elementų kūrimo koeficientas *N* yra parenkamas priklausomai nuo disbalanso duomenyse. Kiekvienam $x \in A$, *N* įrašų yra atsitiktinai parenkami iš artimiausių kaimynų – taip sudaroma aibė A_1.
3. Kiekvienam $x_k \in A_1, (\check{c}ia\ k = 1,2,3,..N)$ yra sukuriamas naujas stebėjimas, remiantis formule:

$$x' = x + rand(0,1) * |x - x_k|, \quad (1)$$

   čia x' - naujas stebėjimas; *rand(0,1)* - atsitiktinis skaičius tarp 0 ir 1.

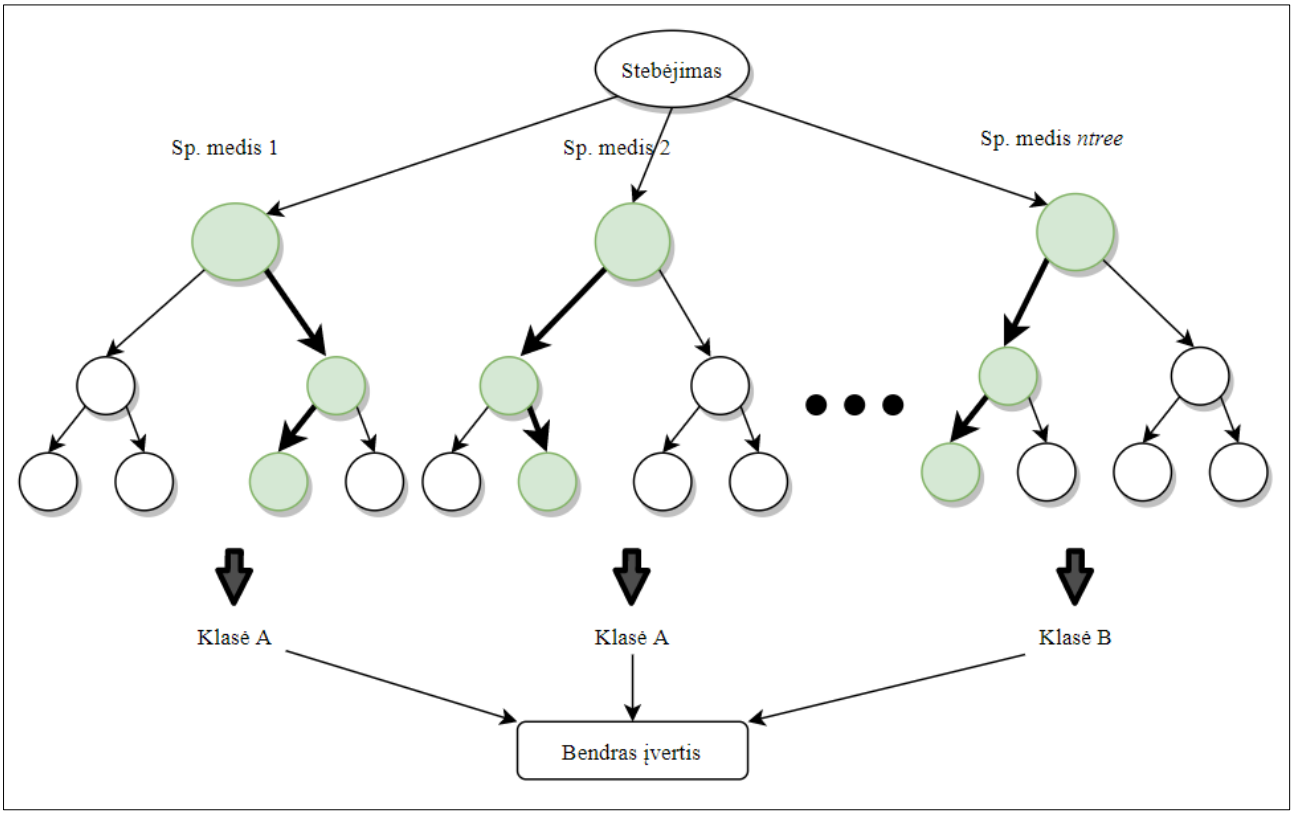
Tyrimo pradžioje, siekiant palyginti modelius, disbalansas yra fiksuojamas. Pradžioje vienai įmonei, kuriai buvo pradėtas bankroto procesas, pasirenkama 10 veikiančių įmonių – naudojamas atsitiktinio mažinimo metodas. Tolesnėje tyrimo eigoje, siekiant nustatyti disbalanso poveikį apmokymo imtyje klasifikavimo rezultatui, bus varijuojamas disbalanso santykis.

2.2. Bankroto prognozavimo modeliai

Šiame skyriuje apžvelgiami bankroto procedūros prognozavimo metodai.

2.2.1. Atsitiktiniai miškai

Atsitiktinių miškų klasifikavimo modelis yra kelių klasifikatorių rinkinys. Atsitiktinį mišką sudarančių medžių skaičius darbe bus žymimas *ntree*. Kiekvienam medžiui iš pradinės duomenų imties yra priskiriama atsitiktinė dalis imties ir atsitiktinai parinkti kintamieji, kurių skaičius bus žymimas *mtry*. Sudaryti medžiai apsimoko ir klasifikuoja, o klasifikavimo kokybei įvertinti naudojami apmokyme nenaudoti šalutiniai OOB (angl. *OOB – out of bag*) duomenys. Atsitiktinio miško klasifikavimo rezultatas yra visų medžių bendras įvertis, atsižvelgiant į kaip dažnai kurią klasę pasirinko medžiai (žr. 8 pav.).



8 pav. Atsitiktinių miškų veikimo schema

Konstruojant sprendimų medžius, kiekvienoje atšakoje yra parenkamas kintamasis iš *mtry* atsitiktinai parinktų kintamųjų. Kintamojo parinkimui yra naudojamas Gini indeksas, kuris nusako sprendimų medžiui priskirtos imties klasių ne grynumą (angl. *impurity*). Gini indeksas apskaičiuojamas įvertinant klasės elemento pasirinkimo tikimybę ir tikimybę, kad jis bus suklasifikuotas neteisingai:

$$Gini = \sum_{i=1}^{c} p_i * (1 - p_i) = 1 - \sum_{i=1}^{c} (p_i)^2 \qquad (2)$$

čia p_i – klasės santykis nagrinėjamoje duomenų imtyje; c – klasių kiekis.

Kadangi atsitiktinių miškų rezultatas yra tikimybė (įvertis nuo 0 iki 1), norint priskirti klasę, reikia parinkti tinkamą slenkstį – nuo kokios reikšmės tarti, kad modelis stebėjimą priskyrė teigiamai klasei ir iki kokios reikšmės – neigiamai. Įprastai praktikoje yra svarbu slenkstį parinkti taip, kad kuo mažiau neigiamos klasės atvejų būtų priskirta teigiamai klasei ir kad būtų atpažinta kuo daugiau teigiamos klasės atvejų.

Naudojant atsitiktinių miškų klasifikavimo metodą, taip pat nustatoma kiekvieno kintamojo svarba klasifikuojant. Kiekviename medžio atsišakojime naudojant atsitiktinai parinktą kintamąjį, nagrinėjama medžiui priskirta duomenų imtis yra padalinama taip, kad tikslo klasės būtų kuo labiau atskirtos. Apmokant medį galima apskaičiuoti, kiek kiekvienas naudojamas kintamasis medžio atsišakojime sumažina klasių ne grynumą. Kuo kintamasis daugiau sumažina ne grynumą, tuo jis yra geresnis. Galutinė kintamųjų svarba yra įvertinama apskaičiuojant visuose medžiuose panaudoto kintamojo sumažinto ne grynumo sumą.

2.2.2. Altmano Z įverčio modelis

Norint gauti koeficientinį įvertį ar įmonei bus pradėtas bankroto procesas ar ne, gali būti naudojamas Altmano Z-įvertis. Įverčio išvedimas yra grįstas santykiniais ekonominiais rodikliais, kurie nusako įmonės finansinį stabilumą:

$$Z = 1.2 * X_1 + 1.4 * X_2 + 3.3 * X_3 + 0.6 * X_4 + 0.99 * X_5 \quad (3)$$

čia X_1 – apyvartinio kapitalo ir turto santykis; X_2 – pardavimo pajamų ir turto santykis; X_3 – pelno neatskaičius mokesčių ir turto santykis; X_4 – nuosavo kapitalo ir įsipareigojimų santykis; X_5 – pardavimo pajamų ir turto santykis.

Kuo apskaičiuotas Z-įvertis yra didesnis, tuo bankroto tikimybė yra didesnė. Šiuo atveju Z-įvertis gali būti apskaičiuojamas kiekvienai įmonei iš metinių duomenų ir apskaičiuota reikšmė gali būti įtraukta kaip vienas aiškinamųjų kintamųjų, taikant atsitiktinio miško klasifikavimo metodą.

2.3. Klasifikavimo tikslumo vertinimas ir metrikos

Kadangi duomenyse yra stiprus disbalansas, klasifikavimo vertinimui turi būti naudojamos atitinkamos metrikos ir metodai – negali būti naudojamas bendras tikslumas, kadangi jis bus daugiau orientuotas į dominuojančią klasę.

Atlikus klasifikavimą, pirminį rezultatą galima suvesti į maišos matricą, kuri apibendrina tikrąsias klasių reikšmes ir prognozes. Maišos matricą sudaro 4 pagrindiniai atvejai (pavyzdys pateiktas 14 lentelėje):

- teisingai teigiami (angl. TP – true positive) – kuomet teigiamos klasės atvejis yra įvertintas teisingai (įmonės, kurioms yra pradėtas bankroto procesas yra atpažintos);
- klaidingai teigiami (angl. FP – false positive) – teigiamos klasės atvejis yra įvertintas klaidingai (įmonės, kurioms pradėtas bankroto procesas, atpažintos kaip gyvuojančios) – I tipo klaida;
- klaidingai neigiami (angl. FN – false negative) – veikiančios įmonės yra priskiriamos prie įmonių, kurioms pradėta bankroto procedūra – II tipo klaida;
- teisingai neigiami (angl. TN – true negative) – veikiančios įmonės yra priskiriamos prie veikiančių įmonių.

14 lentelė. Maišos matricos šablonas.

Prognozuota klasė	**Tikra klasė**	
	NE (vykdo veiklą)	**TAIP (pradėtas bankroto procesas)**
NE (vykdo veiklą)	TN	FN
TAIP (pradėtas bankroto procesas)	FP	TP

Turint maišos matricą galima išvesti kitus tikslumo matus ir metrikas. Atpažinimo arba jautrumo (angl. *recall, sensitivity*) metrika nusako, kiek iš visų teigiamos klasės atvejų modelis suklasifikavo teisingai. Kuo ši metrika didesnė, tuo modelis geresnis.

$$Jautrumas = \frac{TP}{TP + FN} \quad (4)$$

Specifiškumo (angl. *specificity*) metrika nusako, kiek iš visų neigiamos klasės atvejų modelis suklasifikavo teisingai.

$$Specifiškumas = \frac{TN}{TN + FP} \quad (5)$$

Tikslumo (angl. *precision*) metrika nusako, iš visų suklasifikuotų teigiamos klasės atvejų, kiek iš tikrųjų yra priklausančių teigiamai klasei.

$$Tikslumas = \frac{TP}{TP + FP} \quad (6)$$

Lyginant modelius tarpusavyje yra pravartu naudoti F-matą (angl. *F-measure*). Ši metrika padeda įvertinti jautrumą ir tikslumą vienu metu apskaičiuojant jų harmoninį vidurkį, taip per daug nenukrypstant link dominuojančios klasės.

$$F - matas = \frac{2 * Jautrumas * Tikslumas}{Jautrumas + Tikslumas} \quad (7)$$

Kadangi duomenyse yra disbalansas, tyrime negali būti naudojamas bendras tikslumas – jis vertins modelį per daug optimistiškai dėl dominuojančios klasės. Patikimesnė metrika yra bendras balansuotas tikslumas (angl. *balanced accuracy*) – jautrumo ir specifiškumo vidurkis:

$$Balansuotas\ tikslumas = \frac{Jautrumas + Specifiškumas}{2} \quad (8)$$

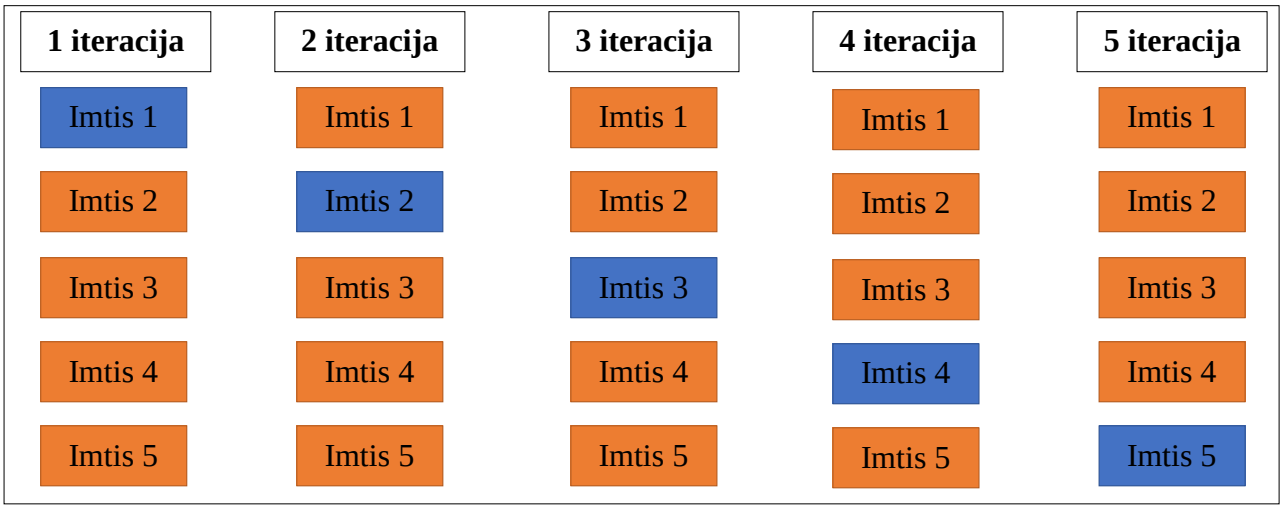
Literatūroje taip pat buvo išskirtas geometrinis vidurkis, kaip tinkamas matas klasifikavimo rezultatui vertinti, kai duomenyse yra disbalansas:

$$Geom.vidurk. = \sqrt{Jautrumas * Specifiškumas} \quad (9)$$

Klasifikavimui vertinti grafiškai bei modelių palyginimui, gali būti naudojamos ROC, DET ir P-R kreivės:

- ROC (angl. *ROC - Receiver Operating Curve*) kreivė – nusako, kaip gerai modelis atskiria klases vieną nuo kitos. X ašyje atidedamos specifiškumo reikšmės nuo 0 iki 1, Y ašyje – jautrumo nuo 0 iki 1. Plotas po kreive yra AUC matas (angl. *AUC – Area Under the Curve*) – kuo šio mato reikšmė didesnė, tuo modelis geriau atskiria įmones, kurioms iškelta bankroto procedūra ir kurioms ne.
- DET (angl. *DET – Detection Error Tradeoff*) kreivė – nusako, kaip blogai modelis klasifikuoja abi klases. X ašyje atidedamos FN, o Y ašyje FP procentinės reikšmės. Iš šios kreivės gaunamas EER matas (angl. *EER – Equal Error Rate*) – tai yra kreivės taškas, kur jautrumas yra lygus specifiškumui. Kuo šis matas mažesnis – tuo modelis geresnis.
- P-R (angl. *Precision-Recall*) kreivė – leidžia vertinti, kaip modelis klasifikuoja tik mažumos klasę. X ašyje atidedamos jautrumo, o Y tikslumo reikšmės nuo 0 iki 1. Kuo ši kreivė labiau išgaubta link taško (1;1), tuo modelis geriau klasifikuoja mažumos klasę.

Taip pat modelių sudarymui ir validavimui tyrime yra naudojamas kryžminis validavimas. Kiekvieną kartą apmokant modelį visa naudojama duomenų imtis yra suskaidoma į *k* nepersidengiančių imčių. Tuomet modelio apmokymui yra naudojama imtis sudaryta iš *k-1* imčių ir viena likusi imtis yra naudojama jo validavimui. Šis procesas yra kartojamas kol visos imčių dalys yra panaudotos tiek apmokyme tiek validavime atskirai, o galutinis rezultatas yra visų iteracijų rezultatų vidurkis. Literatūroje patariama duomenų imtį skaidyti 80% modelio apmokymui ir 20% validavimui, šiuo atveju parametras k bus lygus 5. Kryžminio validavimo schema pateikta **9 pav.**



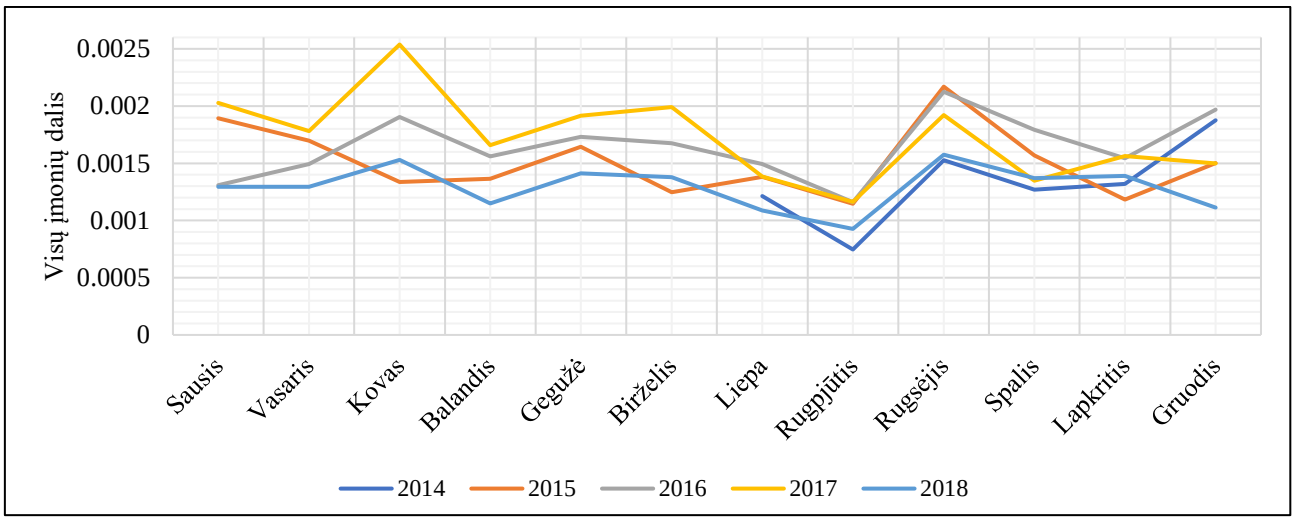
9 pav. Kryžminio validavimo schema, kai k = 5. Oranžinė spalva – imtys, iteracijos metu sudarančios apmokymo duomenų imtį, mėlyna – validavimo duomenų imtis.

3. Tiriamoji dalis

Šiame skyriuje pateikiami atlikti tyrimai ir jų rezultatų analizės.

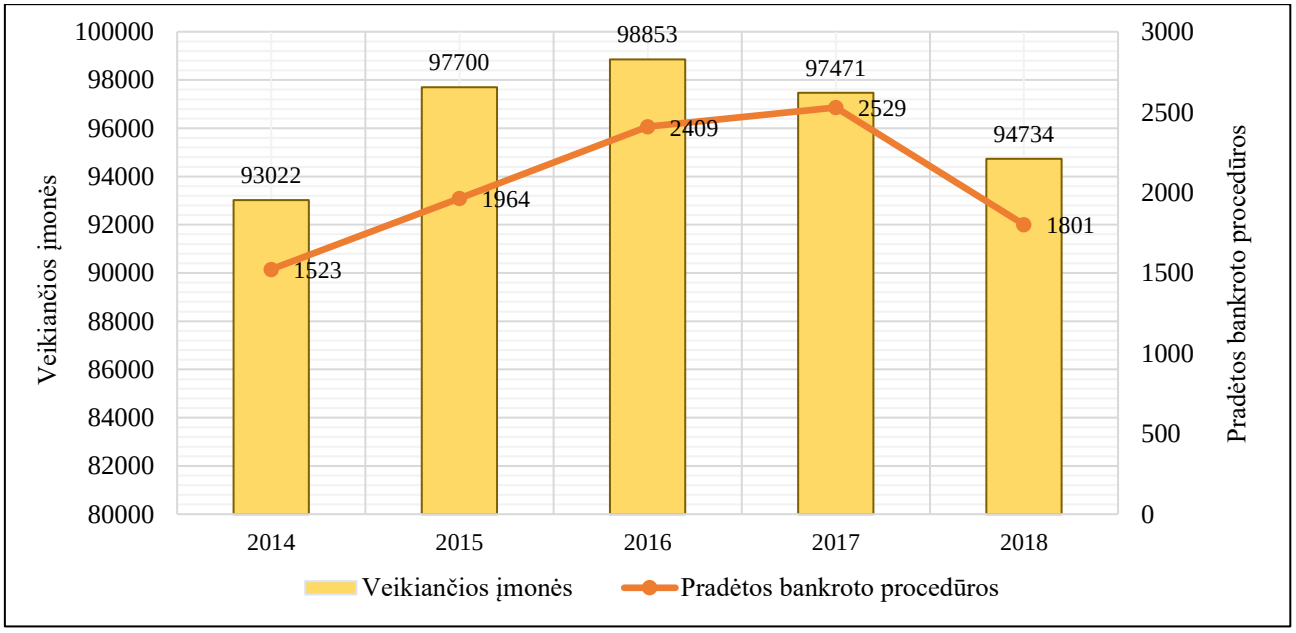
3.1. Duomenų, naudotų tyrime, analizė

Atlikus duomenų valymą ir filtravimą, gaunami galutiniai duomenys, kurie yra toliau naudojami tyrime. Galutines duomenų imtis sudaro 213 kintamieji. Įmonių dalies, kurioms pradedama bankroto procedūra, dinamika pateikta 10 pav..



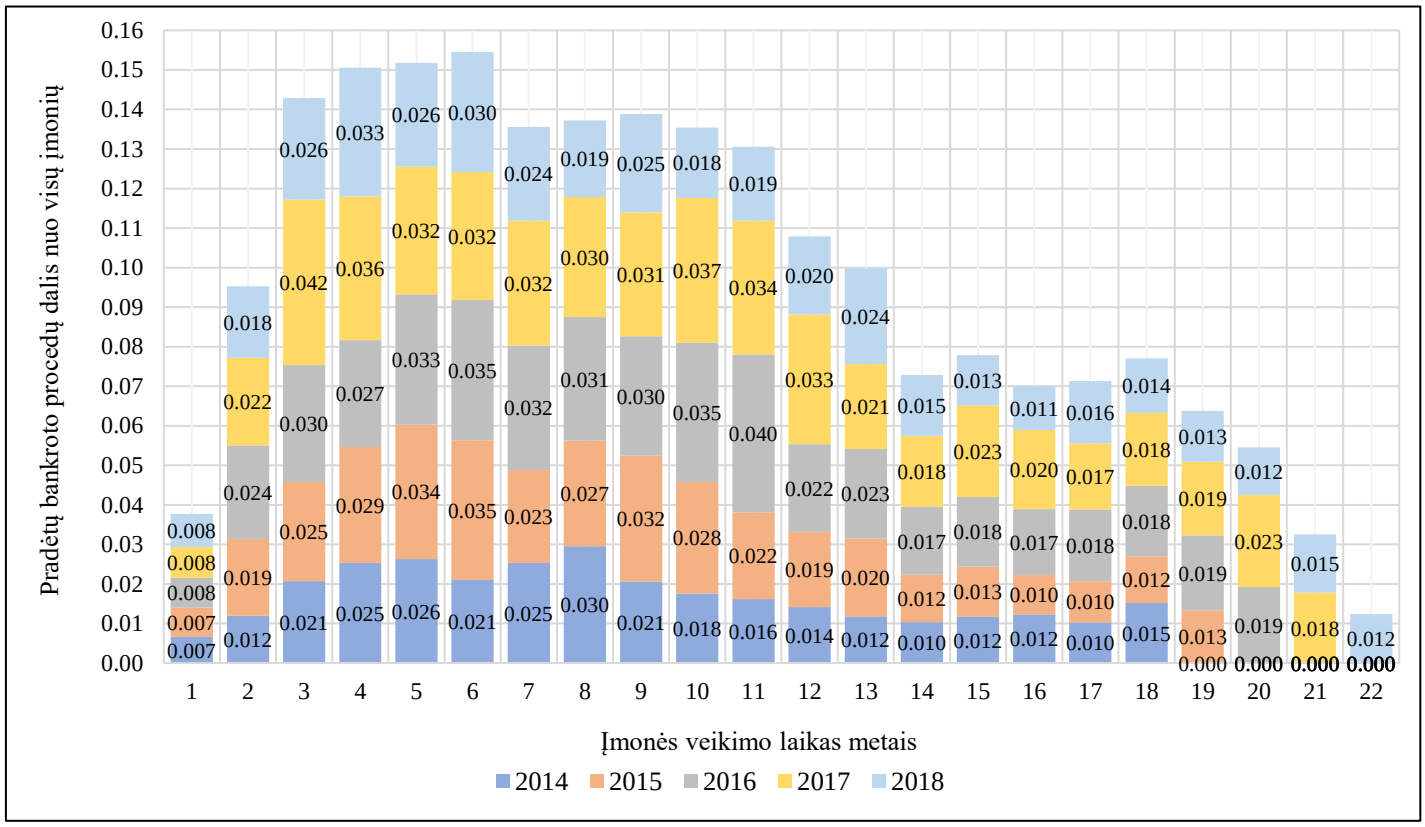
10 pav. Bankrutuojančių įmonių dalis nuo visų veikiančių įmonių pagal mėnesį ir metus

Daugiausiai bankroto procedūrų pradėta 2016, 2017 metais – kovo, rugsėjo mėnesiais. Mažiausiai – 2014, 2018 metais – rugpjūtį. Kadangi 2014 m. įmonės nagrinėjamos tik nuo liepos (kadangi bankrutavusioms iki liepos reikėtų 2012 metų metinių duomenų), visos įmonės, kurioms buvo pradėta bankroto procedūra iki tol – nenaudojamos. Veikiančių ir bankrutuojančių įmonių kaita per metus pateikta 11 pav..



11 pav. Veikiančios ir bankrutuojančios įmonės per metus

Galima pastebėti, kad iki 2016 metų veikiančių įmonių skaičius augo, tačiau 2015 metais priėmus bankroto procedūros skelbimo supaprastinimą, pradėtų bankroto procedūrų skaičius augo taip pat. Nuo 2016 metų veikiančių įmonių skaičius mažėjo. Veikiančių įmonių skaičiaus mažėjimas galėtų būti paaiškintas padaugėjusiomis bankroto procedūromis. Detalesnė apžvalga pagal įmonių gyvavimo laiką pateikiama 12 pav..



12 pav. Įmonių gyvavimo laikotarpis iki bankroto procedūros pradžios pagal metus

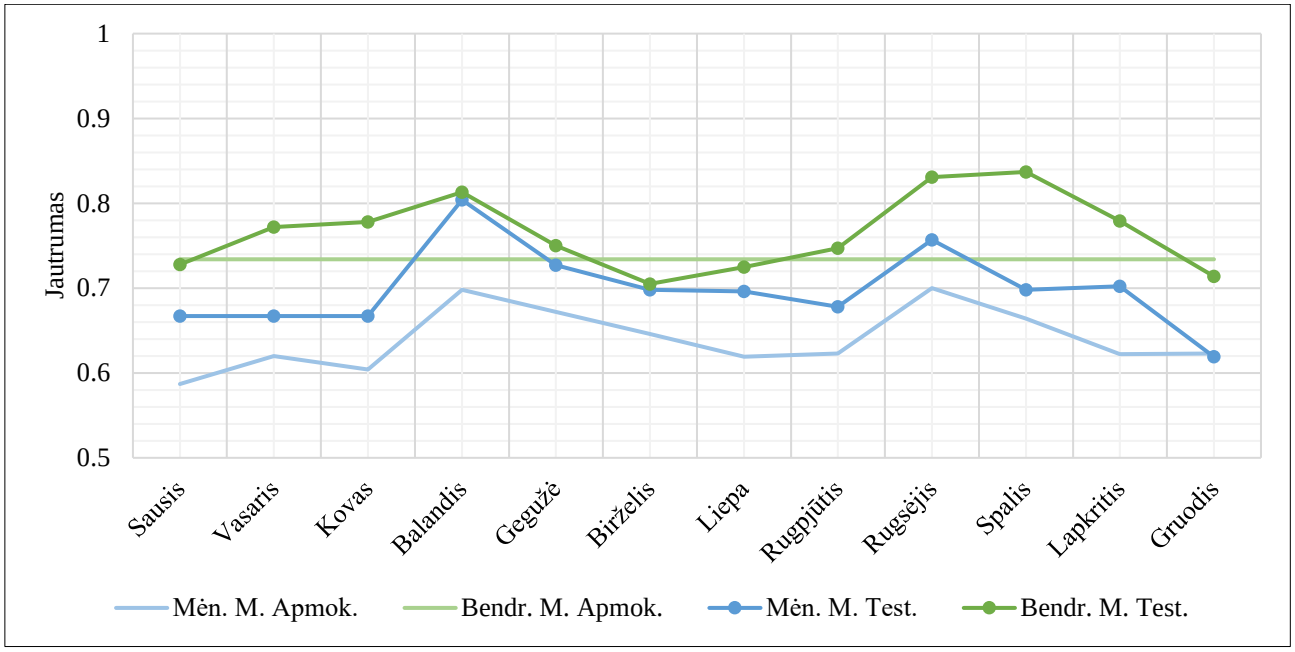
Grafike pateikiama, kokiai daliai visų veikusių įmonių buvo pradėta bankroto procedūra pagal veikimo laiką. Pirmais veikimo metais per visus 2014-2018 metus vidutiniškai, bankroto procedūros buvo pradėtos tik 0.76% visų veikusių įmonių. Kai tuo tarpu antrus metus veikiančioms įmonėms nagrinėjamame periode, bankroto procedūra buvo pradėta vidutiniškai 1.9% įmonių. Galima pastebėti, kad daugiausiai linkusios bankrutuoti naujos įmonės – nuo 3 iki 6 metų amžiaus. Tai atitinka Belgijos mokslininkų tyrimą, kuriame papildomai buvo nustatyta, kad jaunos, dar tik įsitvirtinančios įmonės yra dažniau bankrutuojančios nei kitos [13]. Daugiausiai – vidutiniškai 3.06% visų šeštus metus veikusių įmonių buvo pradėta bankroto procedūra.

Vertinant bankroto situacijos specifiką Lietuvoje, dėl 2015 metais supaprastinto bankroto procedūros skelbimo 2015-2017 metais išaugo pradėtų bankroto procedūrų skaičius ilgai veikusioms įmonėms. Tai atsispindi grafike – 2016-2017 metais 14-19 metus gyvavusioms įmonėms pradėtų bankroto procedūrų skaičius buvo didžiausias per visą nagrinėjamą laikotarpį. Tais metais bankroto procedūros buvo pradėtos 1.8% visų 14-19 metų veikusių įmonių, kai 2014-2015 metais – 1.1%. Tai gali būti viena priežasčių, kodėl apskritai 2016-2017 metais buvo pasiektas didžiausias pradėtų bankroto procedūrų skaičius.

3.2. Metinio ir mėnesinių modelių palyginimas

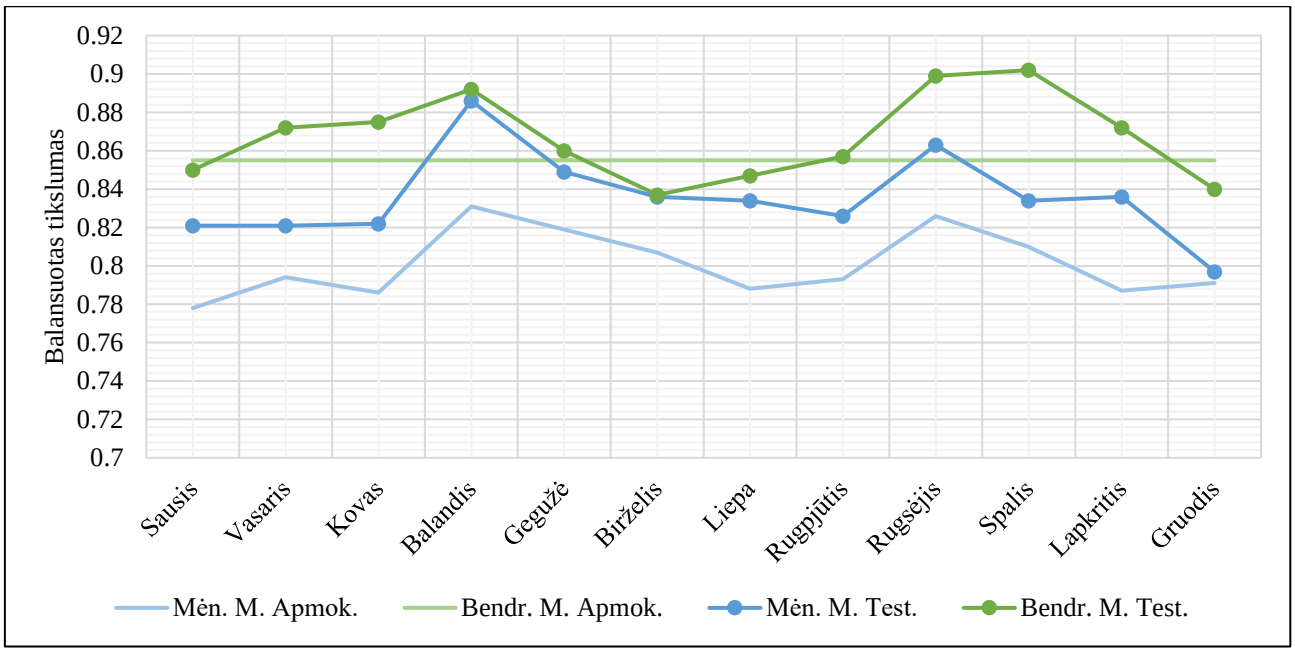
Norint palygint skirtingus modelius, reikia sudaryti vienodas palyginimo sąlygas – fiksuoti kintamuosius modelių sudarymui. Konstruojant atsitiktinių miškų modelius pasirenkami preliminaraus tyrimo metu nustatyti parametrai: medžių skaičius (*ntree*) - 200, atsitiktinai pasirenkamų kintamųjų skaičius šakojant medį (*mtry*) - 50, mažiausias duomenų eilučių skaičius panaudojamas medžio atsišakojime (*min.node.size*) – 100. Šie parametrai nustatyti naudojant automatinį atsitiktinių miškų derinimą - *tuneRanger()* funkciją. Tikimybės slenkstis rezultatui klasifikuoti taip pat pasirenkamas bendras, atsižvelgiant į rekomenduojamus kiekvienam atvejui - 0.35. Naudojant kryžminį validavimą, imtis yra suskaidoma į 5 dalis (80% apmokymui, 20% validavimui).

Modelių kūrimui yra naudojami 213 kintamųjų. Iš viso sukuriama 13 modelių – 1 apmokytas visų mėnesių duomenimis ir 12 apmokytų tik specifinių mėnesių duomenimis. Apmokymui naudojamų duomenų laikotarpis yra nuo 2013 iki 2017 metų. Testavimui naudojami 2018 metų kiekvieno mėnesio duomenys. Modeliai lyginami pagal pagrindines metrikas: jautrumą, specifiškumą, F1, balansuotą tikslumą, AUC ir EER metrikas.



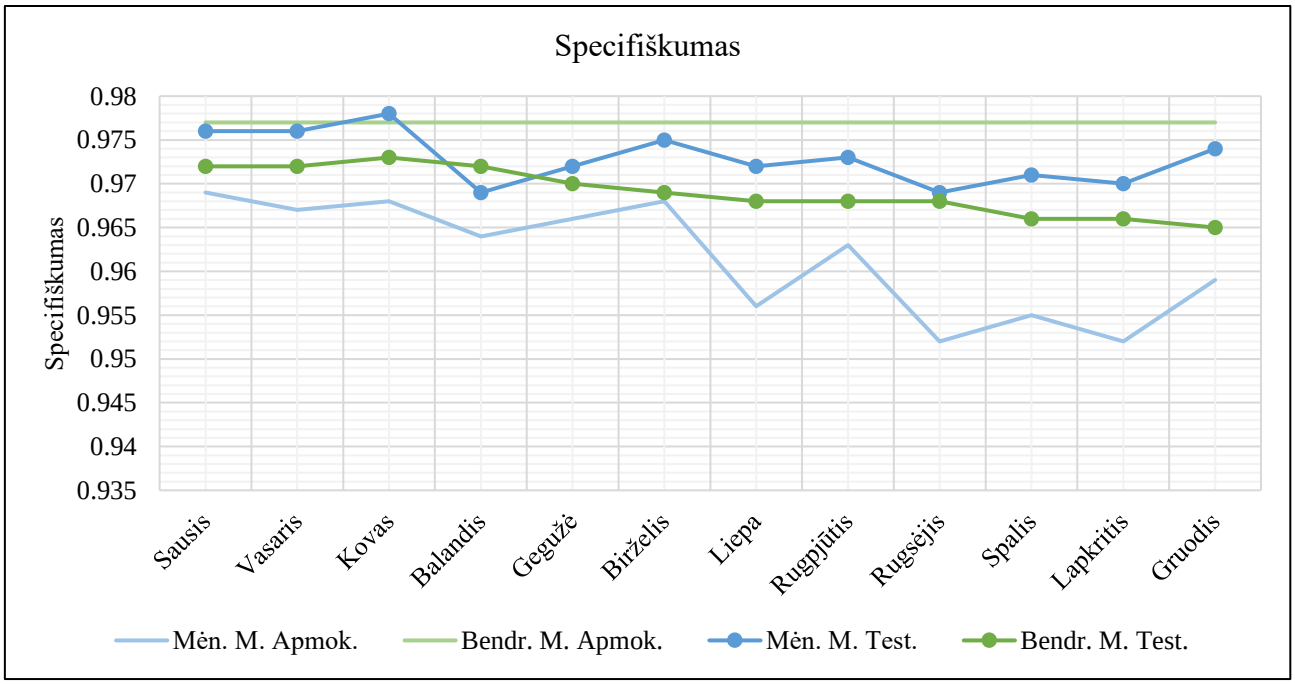
13 pav. Bendro ir mėnesinių modelių palyginimas pagal jautrumo metriką.

Lyginant modelius pagal jautrumo metriką (žr. 13 pav.), atsižvelgiant į testavimo rezultatus, bendras modelis pasiekia aukštesnį jautrumą, testuojant su kiekvieno mėnesio duomenimis – bendras modelis atpažįsta daugiau iš tikrųjų bankrutuojančių įmonių. Tą patį galima pamatyti nagrinėjant balansuoto tikslumo metriką (žr. 14 pav.).



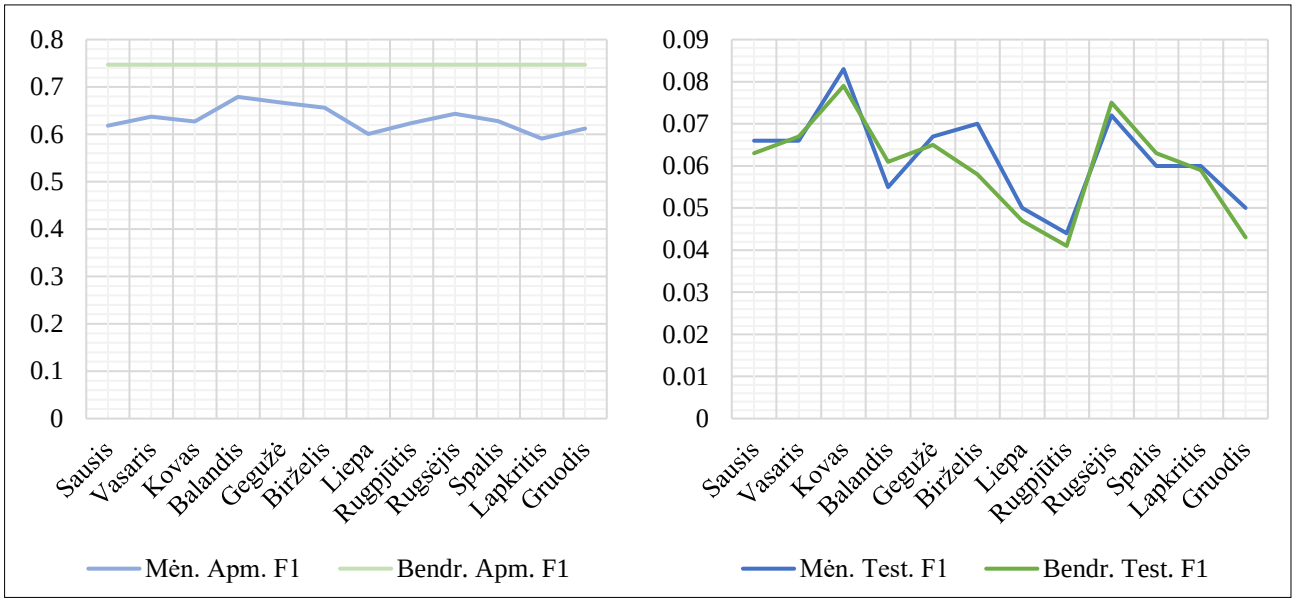
14 pav. Bendro ir mėnesinių modelių palyginimas pagal balansuoto tikslumo metriką.

Kadangi jautrumo ir specifiškumo metrikos yra naudojamos balansuotam tikslumui apskaičiuoti, grafikai yra panašūs ir bendras modelis yra pranašesnis, vertinant šias metrikas bendrai. Tačiau lyginant modelius pagal specifiškumą atskirai, gaunami kitokie rezultatai (žr. 15 pav.).



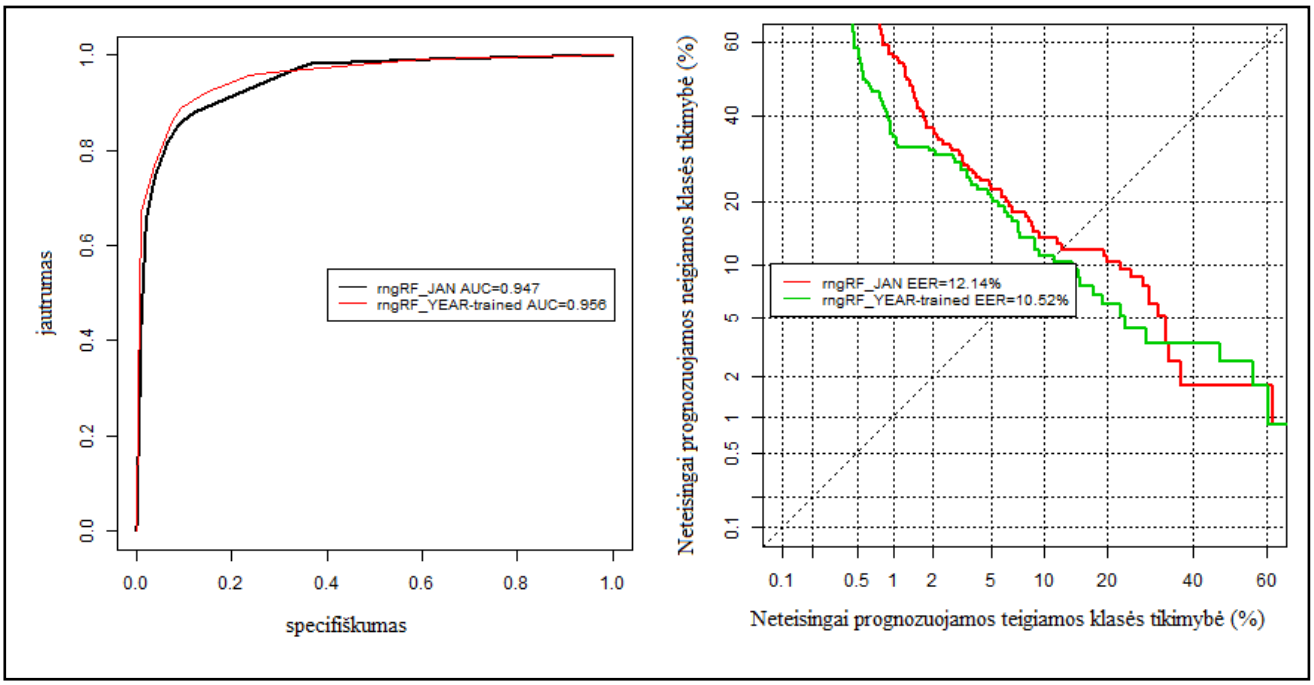
15 pav. Bendro ir mėnesinių modelių palyginimas pagal specifiškumo metriką.

Atsižvelgiant į specifiškumo metriką, beveik visais atvejais mėnesiniai modeliai turėjo aukštesnę reikšmę – būtų galima sakyti, kad mėnesinis modelis geriau atskiria veikiančias įmones, tačiau didžiausias skirtumas tarp mėnesinio ir bendro modelio nesiekė 0.01, todėl rezultatai labai panašūs.



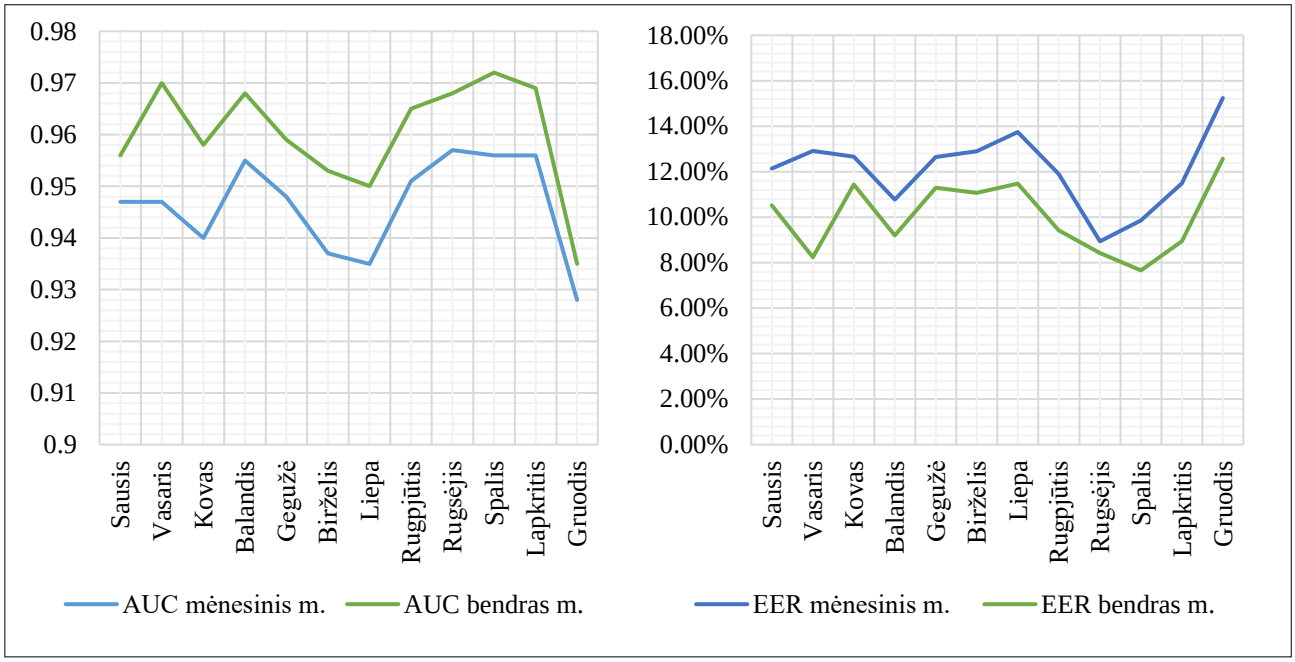
16 pav. Bendro ir mėnesinių modelių palyginimas pagal F1 metriką

F1 metrika apima jautrumo ir tikslumo metrikas. Šiuo atveju jau yra vertinama ir kiek iš visų bankrutuojančioms priskirtų įmonių iš tiesų yra bankrutuojančios. Ši metrika abiejų tipų modeliams stipriai skiriasi, lyginant tarp apmokymo ir testavimo imčių. Tai gali būti paaiškinta skirtingu klasių disbalansu apmokymo ir testavimo imtyse (testavimo imtyje klasių disbalansas yra didesnis). F1 metrika testavimo imtyje tarp bendro ir mėnesinių modelių daug nesiskiria – 5 mėnesiais iš 12 bendras modelis prognozuoja geriau, tačiau skirtumas nėra didelis – iki 0.01. Toliau modeliai lyginami pagal AUC ir EER metrikas. Pateikiamas testavimo su sausio duomenimis rezultato AUC ir EER metrikų fragmentas (žr. 17 pav.).



17 pav. Abiejų modelių sausio mėnesio rezultatų ROC ir DET kreivės su AUC ir EER reikšmėmis.

Galima pastebėti, kad šiuo atveju abi metrikos, naudojant bendrą modelį, yra geresnės nei taikant specifinį sausio mėnesio modelį. Geresnis modelis turi aukštesnę AUC ir mažesnę EER reikšmes. Apibendrinti AUC ir EER metrikų rezultatai visų mėnesių rezultatams pateikiami 18 pav.eiksle.



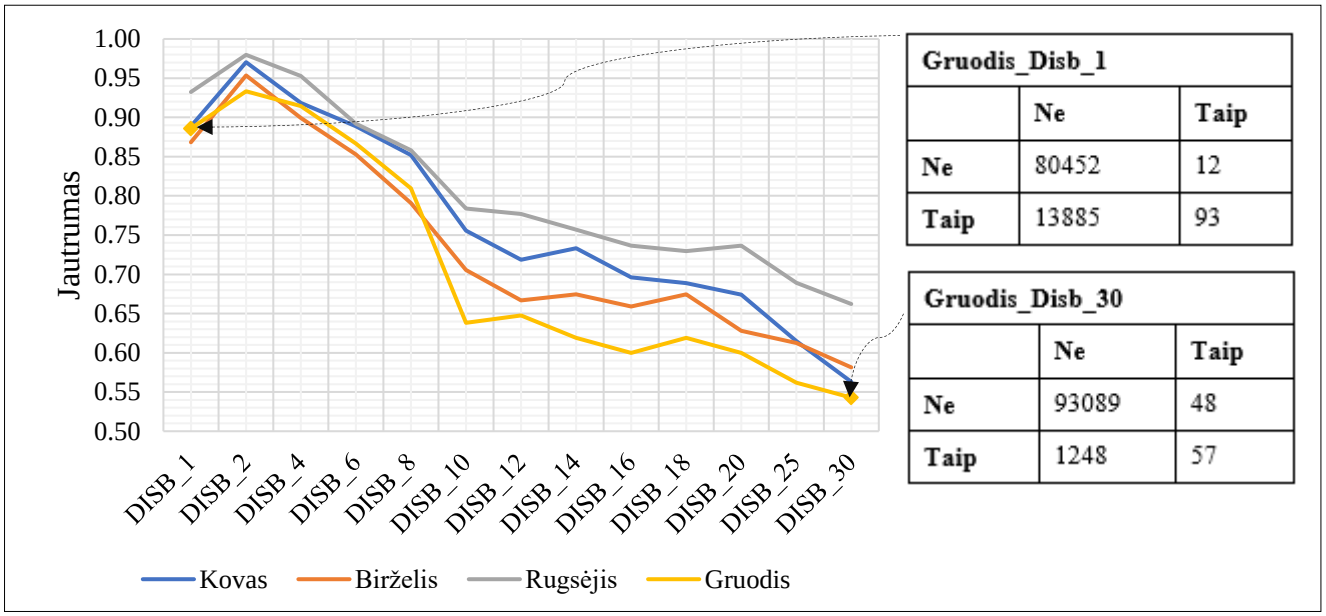
18 pav. Bendro ir mėnesinių modelių palyginimas pagal AUC ir EER metrikas

Tiek pagal AUC metriką, tiek pagal EER – bendras modelis yra pranašesnis prognozuojant visais mėnesiais. Nagrinėjant AUC metriką – modelio rezultatą vertinant su įvairias slenksčiais, visais mėnesiais bendras modelis yra geresnis. Pagal EER metriką, vidutiniškai mažiausia tikimybė veikiančią įmonę priskirti bankrutuojančioms ir bankrutuojančią – veikiančioms yra pasiekiama taikant bendrą modelį.

Galima daryti išvadą, kad taikant bendrą visų mėnesių modelį - apmokymui naudojant visų mėnesių duomenis, pasiekiamas geresnis rezultatas nei naudojant tik tam tikro mėnesio duomenis. Taip pat modelio apmokymui naudojant nuo fiksuotos datos pernai ar užpernai deklaruotus metinius duomenis, rezultatas nuo to reikšmingai nesikeičia. Viena priežastis gali būti, kad skirtingų mėnesių prognozei naudojant tų pačių metų metinius duomenis, vėlavimas iki metinių duomenų bus skirtingas – sausio mėnesiui skirtumas 10 mėnesių (pirmas naudojamas mėnesinis įrašas spalio mėnesį, iki užpraeitų metų gruodžio - 10 mėnesių), o vasario - 11 mėnesių. Taip pat įmonių, skaičius, kurioms buvo pradėta bankroto procedūra kiekvieną mėnesį skiriasi, tai taip pat daro įtaką modelių klasifikavimo rezultatams, vertinant pagal mėnesius.

3.3. Disbalanso duomenyse įtaka bankroto prognozei

Siekiant nustatyti, kaip kinta galutinis klasifikavimo rezultatas, kintant disbalansui apmokymo duomenyse, iš pradžių iš pradinių duomenų išskiriama 13 imčių – kiekvienoje imtyje yra visos įmonės, kurioms pradėtas bankroto procesas (7704) ir atitinkamas skaičius veikiančių įmonių. Veikiančių įmonių skaičius vienai bankrutuojančiai svyruoja nuo 1 iki 30. Apmokius 13 modelių kiekvienam duomenų disbalansui, modeliai yra testuojami su kiekvieno mėnesio duomenimis.



Gruodis_Disb_1		
	Ne	Taip
Ne	80452	12
Taip	13885	93

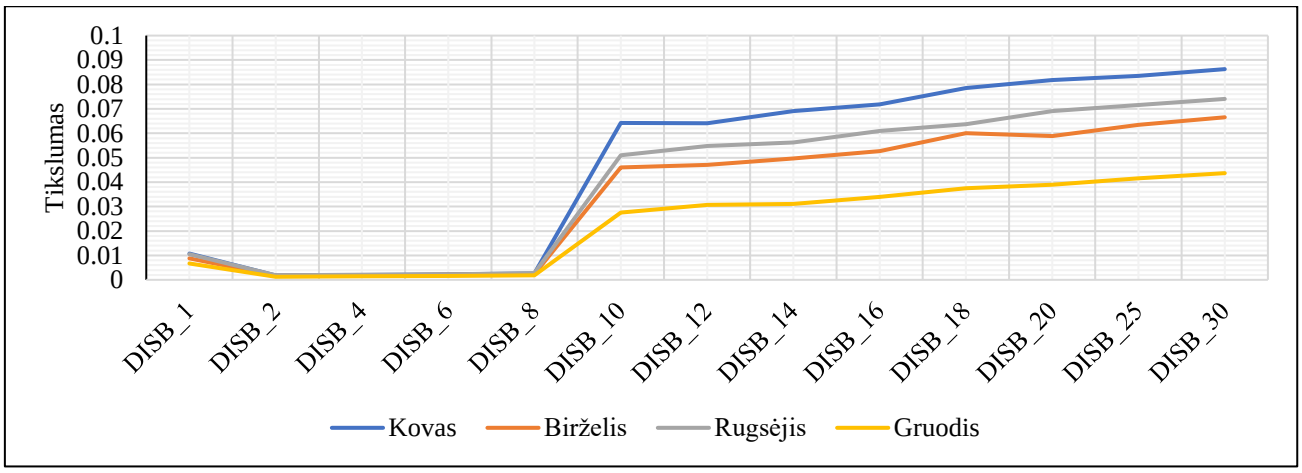
Gruodis_Disb_30		
	Ne	Taip
Ne	93089	48
Taip	1248	57

19 pav. Modelių, apmokytų su skirtingo disbalanso duomenimis, palyginimas testavimo imtyse pagal jautrumą

Lyginant modelius, sukurtus su skirtingu disbalansu apmokymo imtyje, pagal jautrumo metriką (žr. 19 pav.) gaunamas rezultatas, kokio ir gali būti tikimasi. Modeliai, kurių apmokymo imtyje disbalansas buvo mažas (1,2,4,...), geriau atpažįsta įmones, kurioms gali būti pradėta bankroto procedūra, tačiau labai daug veikiančių įmonių yra įtraukiama į rizikingų sąrašą – tai nėra optimalus sprendimas, vertinant iš praktinės pusės. Modeliai, kuriems apmokyti buvo įtraukta daugiau veikiančių įmonių – apmokymo duomenyse disbalansas buvo didesnis – beveik 10 kartų mažiau veikiančias įmones priskiria bankrutuojančioms, tačiau prasčiau skiria iš tikrųjų bankrutuojančias.

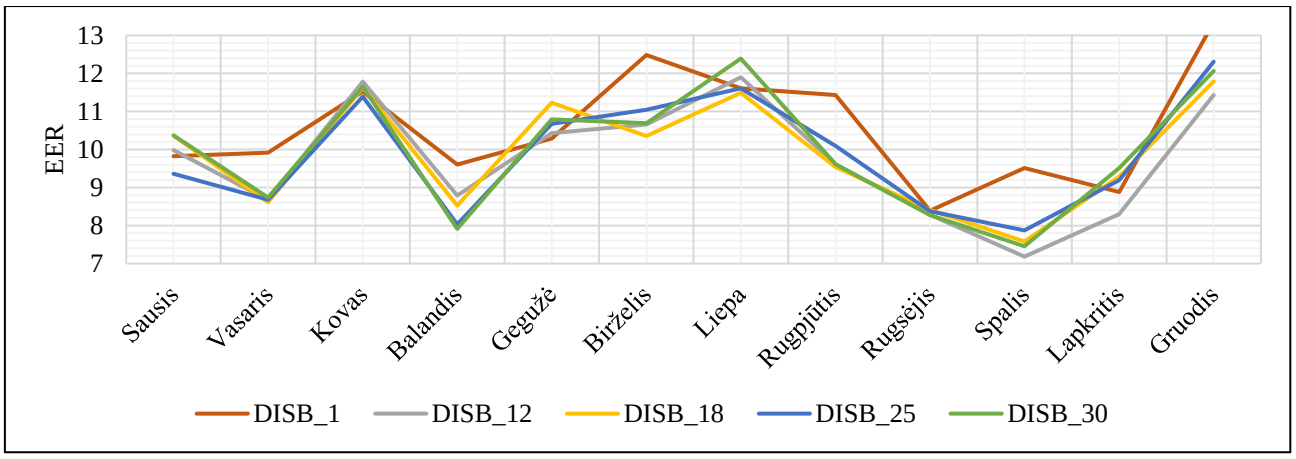
Kadangi jautrumas nusako tik kaip gerai yra klasifikuojama teigiama klasė, į neigiamą klasę ši metrika neatsižvelgia, natūralu, kad vertė yra didžiausia, kai disbalansas apmokymo duomenyse yra mažiausias.

Nagrinėjant skirtumą tarp mėnesinių testavimo duomenų, galima pastebėti, kad modelis bankrutuojančias įmones rugsėjui prognozuoja geriau nei gruodžiui. Taip gali būti dėl to, nes rugsėjį apskritai yra daugiau įmonių, kurioms pradėta bankroto procedūra – modelis buvo geriau apmokytas.

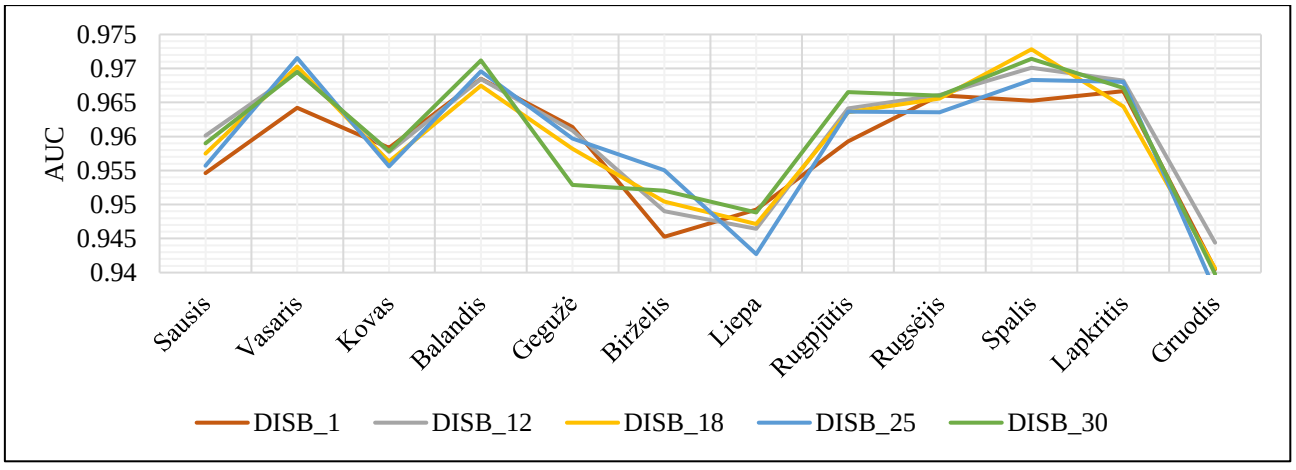


20 pav. Modelių, apmokytų su skirtingo disbalanso duomenimis, palyginimas testavimo imtyse pagal tikslumą

Atsižvelgiant į tikslumo metriką, įvertinama ir dalis įmonių, kurios yra veikiančios, tačiau buvo priskirtos prie bankrutuojančių. Šiuo atveju, kadangi mažo disbalanso duomenimis apmokyti modeliai apskritai nemažai veikiančių įmonių priskiria bankrutuojančioms, o iš tikrųjų bankrutuojančių yra mažai - gaunama, kad ši metrika yra itin maža.



21 pav. Modelių, apmokytų su skirtingo disbalanso duomenimis, palyginimas testavimo imtyse pagal EER

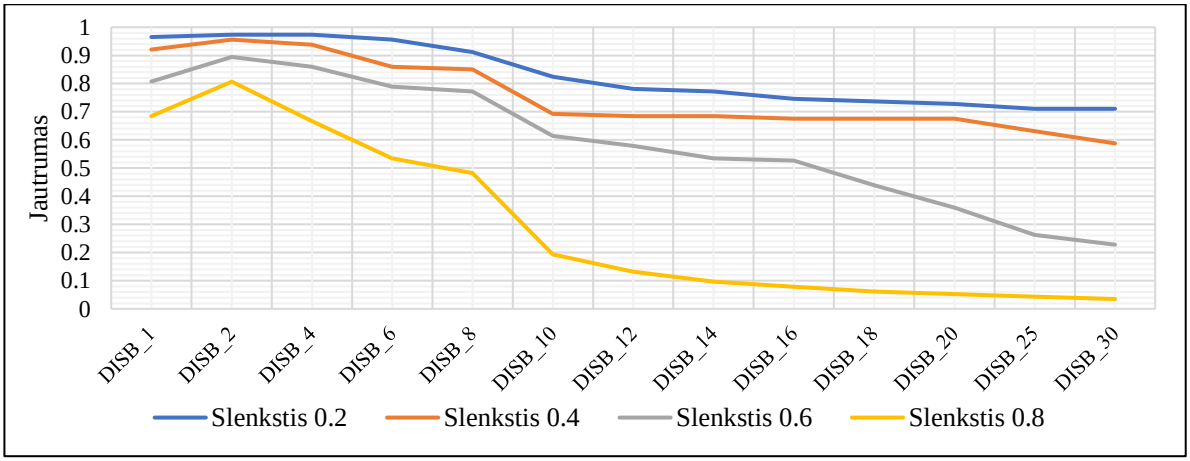


22 pav. Modelių, apmokytų su skirtingo disbalanso duomenimis, palyginimas testavimo imtyse pagal AUC

Pagal AUC ir EER metrikas modeliai klasifikuoja panašiai. EER metrika didžiausia modeliui, kuris apmokytas su mažiausiu disbalansu, atitinkamai šiuo modeliu buvo pasiekta vidutiniškai mažiausia AUC reikšmė.

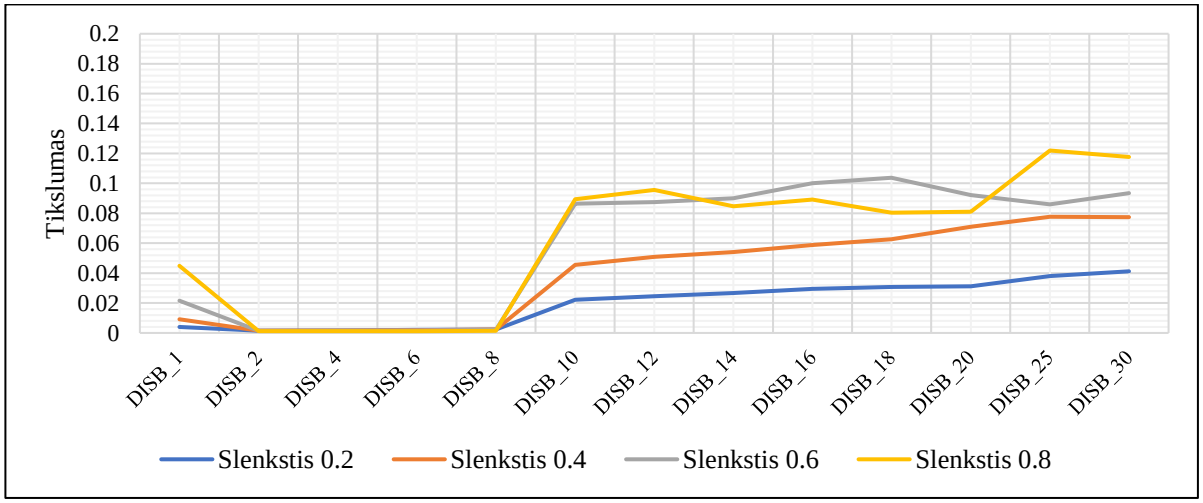
3.4. Rekomenduojama slenksčio vertė

Priklausomai nuo išsikelto tikslo klasifikuojant yra svarbu, kokia slenksčio reikšmė yra pasirenkama. Jeigu svarbu atskirti kuo daugiau įmonių, kurioms gali būti pradėta bankroto procedūra ir nėra didelės svarbos, kiek veikiančių įmonių yra priskiriama prie bankrutuojančių, atsižvelgiant į klasifikavimo metrikų specifikas, slenksčio parinkimą galima vykdyti pagal kuo didesnę jautrumo metriką (žr. 23 pav.).



23 pav. Sausio mėnesio klasifikavimo rezultatas pagal jautrumo metriką, priklausomai nuo disbalanso apmokymo imtyje ir slenksčio dydžio

Galima pastebėti, kad šiuo nagrinėjamu atveju, pasirinkus mažesnę slenksčio reikšmę, geriau atskiriamos iš tikrųjų bankrutuojančios įmonės. Siekiant išskirti tik bankrutuojančias, neteikiant svarbos veikiančių įmonių priskyrimui prie bankrutuojančių, esant dideliam disbalansui, geriau naudoti mažą slenkstį, šiuo atveju – 0.2. Esant mažam disbalansui, slenksčio parinkimas didelės reikšmės neturi.



24 pav. Sausio mėnesio klasifikavimo rezultatas pagal tikslumo metriką, priklausomai nuo disbalanso apmokymo imtyje ir slenksčio dydžio

Jeigu vertinamas tiek bankrutuojančių, tiek veikiančių įmonių klasifikavimo teisingumas, atsižvelgus į tikslumo metriką (žr. 24 pav.), galima matyti, kad mažo slenksčio atveju apskritai daug įmonių yra priskiriama prie bankrutuojančių, iš kurių tik maža dalis yra tikrai bankrutuojančių. Esant mažam disbalansui galima naudoti didesnį slenkstį (0.6 – 0.8), kadangi jautrumo metrikai tai didelės įtakos neturi, o dideliam disbalansui reikėtų naudoti mažesnį – 0.4 slenkstį, kadangi su šia reikšme per daug nenukenčia jautrumo metrika ir tikslumo metrika stipriai nesiskiria nuo tų, kurios pasiekiamos su didesniais slenksčiais.

Svarbios tik bankrutuojančios įmonės

Sausis_Disb_1__Slenks_0.2		
	Ne	**Taip**
Ne	61230	4
Taip	26770	110

Sausis_Disb_1_Slenks_0.6		
	Ne	**Taip**
Ne	83803	22
Taip	4197	92

Sausis_Disb_30__Slenks_0.2		
	Ne	**Taip**
Ne	86115	33
Taip	1885	81

Svarbios ir bankrutuojančios ir veikiančios įmonės

Sausis_Disb_1_Slenks_0.6		
	Ne	**Taip**
Ne	83803	22
Taip	4197	92

Sausis_Disb_1__Slenks_0.8		
	Ne	**Taip**
Ne	86340	36
Taip	1660	78

Sausis_Disb_30_Slenks_0.4		
	Ne	**Taip**
Ne	87201	47
Taip	799	67

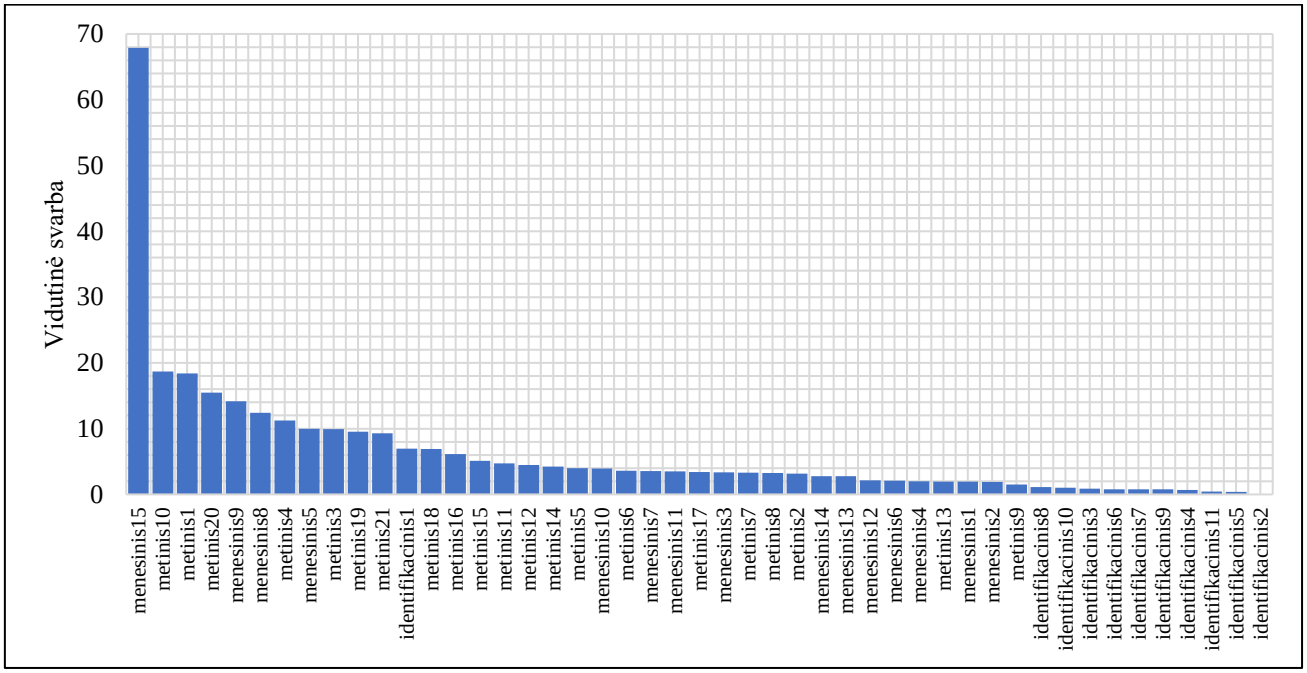
25 pav. Slenksčio parinkimo tyrimo apibendrinimas

Slenksčio parinkimo tyrimo apibendrinimas pateikiamas 25 pav**.**eiksle. Galima pastebėti, kad parinkus atitinkamas slenksčių vertes (0.8 ir 0.4), modelis, apmokytas su didelio disbalanso duomenimis, klasifikuoja panašiai, kaip modelis, apmokytas su mažo disbalanso duomenimis. Vienu atveju geriau išskiriamos veikiančios įmonės, kitu – bankrutuojančios.

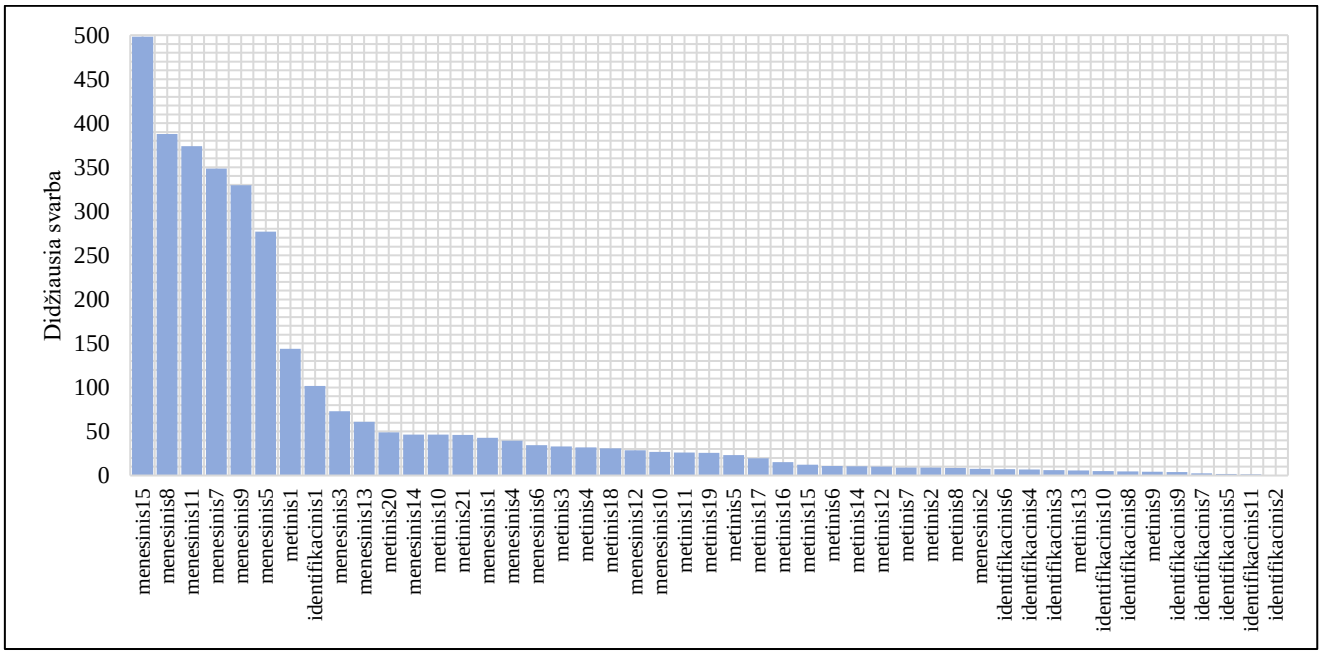
Reguliuojant slenksčio parinkimą, bankroto procedūros prognozavimo modeliai gali būti pritaikyti įvairioms praktikoms. Kreditų davėjams, siekiant nustatyti ar įmonė ateityje gebės įvykdyti savo įsipareigojimus, bendru atveju nėra būtina išskirti tik tas įmones, kurioms gali būti pradėta bankroto procedūra. Veikiančios įmonės, modelio priskirtos prie bankrutuojančių, taip pat gali turėti indikatorių apie finansinius sunkumus tik, galbūt, ne tokius stiprius, kad ateinantį mėnesį įmonei jau būtų pradėta bankroto procedūra. Tokiu atveju įmonių priskyrimas bankrutuojančioms gali būti reikšmingas ir, nepriklausomai nuo disbalanso, parinkus mažą slenksčio vertę būtų atmesta didelė dalis iš tikrųjų bankrutuojančių įmonių ir dalis įmonių, kurios susiduria su finansiniais sunkumais. Pavyzdžiui, vertinant iš mokesčių inspekcijos perspektyvos - norima kuo anksčiau pastebėti įmones, kurioms ateityje gali būti pradėta bankroto procedūra, siekiant anksčiau užšaldyti įmonės sąskaitas ir areštuoti turtą, kol įmonė jo neišpardavė. Tokiu atveju praktiškiau būtų mažas bendras sąrašas įmonių, jeigu kiekviena galimai probleminė įmonė yra papildomai patikrinama, kas reikalauja papildomų resursų.

3.5. Reikšmingiausi kintamieji

Atlikus bandymus ir surinkus rezultatus iš visų tyrimuose taikytų modelių, atliekama bendra statistinė analizė, siekiant nustatyti reikšmingiausius kintamuosius (žr. 26 pav. ir 27 pav.).



26 pav. Vidutinė kintamųjų svarba



27 pav. Didžiausia kintamųjų svarba

Galima pastebėti, kad svarbiausi kintamieji, prognozuojant bankrotą buvo likutis – nepriemoka VMI (*mėnesinis15*), toliau, kaip ir išskirta literatūroje – nuosavas kapitalas (*metinis10*) bei grynasis pelnas (*metinis4*). Anksčiau atliktoje duomenų apžvalgoje nustatytą įmonės amžiaus ir bankroto ryšį (taip pat pateikiamą ir literatūroje) atitinka reikšmingas kintamasis - įmonės amžius (*metinis1*).

Literatūroje nenagrinėti duomenys apie darbuotojus – darbo užmokesčio kaštai (*mėnesinis9*), viso darbuotojų (*mėnesinis7*) yra vieni svarbiausių, prognozuojant bankroto pradžios procedūrą.

Patys nereikšmingiausi kintamieji yra identifikaciniai - kategoriniai, paversti į binarinius kintamuosius. Bankroto pradžios procedūros prognozei reikšmės neturi įmonės ekonominės veiklos rūšis, taip pat įmonės lokacija. Vienintelis naudingas indikatorius yra ar įmonė yra mažasis mokesčių mokėtojas (*identifikacinis1*).

Remiantis šiais rezultatais, galima daryti išvadą, jog vienas pagrindinių bankroto indikatorių yra didelės arba augančios įmonės skolos VMI. Užsienio literatūroje panašūs reikšmingiausi kintamieji buvo trumpalaikiai ir ilgalaikiai įsipareigojimai, šiuo atveju, skola VMI yra konkretesnis ir svarbesnis indikatorius – šis rodiklis yra mėnesinis, kas stipriai padeda stebėti įmonės būseną trumpuoju periodu. Taip pat, jeigu įmonės metinis grynasis pelnas yra mažas ar neigiamas, tai taip pat gali būti finansinių sunkumų indikatorius. Iš turimų duomenų: 36 % visų veikiančių įmonių deklaravo neigiamą metinį grynąjį pelną, tuo tarpu visų bankrutuojančių – 71%. Nagrinėjant įmonės amžių, atliktoje duomenų žvalgomojoje analizėje, galima pastebėti, kad įmonės, gyvavusios 3-6 metus yra labiau linkusios bankrutuoti nei ilgai veikiančios – tai atitinka literatūros analizėje aprašytą Belgijos mokslininkų išvadą. Detaliau apžvelgus nuosavą kapitalą – veikiančiose įmonėse 21% įmonių deklaravo neigiamą nuosavą kapitalą, bankrutuojančių įmonių – 58%. Užsienio tyrimuose nebuvo nagrinėjami duomenys apie darbuotojus, tačiau šiuo atveju kai kurie rodikliai apie darbuotojus išskiriami kaip vieni reikšmingiausių. Veikiančios įmonės per metus darbuotojų skaičių padidino daugiau nei 8%, bankrutuojančios per metus darbuotojų skaičių sumažino daugiau nei 27%. Įmonės, kurios turi finansinių sunkumų, siekiant sutaupyti, laikui einant pradeda atleidinėti darbuotojus. Verta pabrėžti, kad nagrinėjant veikiančias įmones, kurios deklaruoja prastus finansinius rodiklius ar yra modelio priskirtos prie bankrutuojančių, nėra vertinama ar joms buvo pradėta bankroto procedūra už didesnio laiko periodo. Kaip literatūroje pabrėžiama – bankrotas yra laike besitęsiantis procesas, kurio trukmė gali labai įvairiai svyruoti.

Išvados

1. Apžvelgus literatūrą pastebėta, kad bankroto prognozės problema yra plačiai nagrinėjama įvairiais aspektais. Sprendžiant šią problemą, susiduriama su įvairiais papildomais uždaviniais – disbalanso sprendimu, kintamųjų parinkimu, prognozės metodo parinkimu, kuriems spręsti naudojami mašininio mokymo algoritmai (atsitiktinis miškas, neuroniniai tinklai, logistinė regresija), duomenų disbalanso tvarkymo metodai (atsitiktinis išmetimas, mažumos klasės dauginimas), kintamųjų parinkimo metodai (atsitiktinis miškas, Lasso metodas, remiantis literatūra). Atsižvelgiant į literatūros analizę, šiame tyrime pasirinkta naudoti atsitiktinio miško bei atsitiktinio išmetimo algoritmus.
2. Išanalizavus gautus duomenis, pastebėta, kad per mėnesį vidutiniškai tik 0.2% visų veikiančių įmonių buvo pradėta bankroto procedūra. Daugiausiai bankroto procedūrų pradėta 3-6 metus veikiančioms įmonėms, mažiau – ilgai veikiančioms įmonėms. Duomenys buvo konvertuoti į mašininio mokymo algoritmams tinkamą formatą.
3. Sukūrus ir palyginus bendrą visų mėnesių ir skirtingus specifinių mėnesių modelius, nustatyta, kad bendras modelis, kurio apmokymui buvo naudoti visų metų mėnesių duomenys prognozavo tiksliau. Galima daryti išvadą, kad metinių rodiklių vėlavimo skirtumas apmokymo duomenyse neturi reikšmės bankroto prognozei.
4. Sukūrus modelius su įvairiu disbalansu apmokymo imtyje, nustatyta, kad esant didesniam disbalansui, geriau atskiriamos veikiančios įmonės, bet prasčiau bankrutuojančios. Kuo disbalansas apmokymo duomenyse mažesnis, tuo geriau atskiriamos bankrutuojančios įmonės – stiprus rezultatų pagerėjimas tikslumo atžvilgiu pasiekiamas kai apmokymo imtyje yra 10 ir daugiau veikiančių įmonių 1 bankrutuojančiai.
5. Praktikoje naudingesniu atveju – siekiant teisingai atskirti kuo daugiau bankrutuojančių įmonių ir kuo mažiau veikiančių įmonių priskirti bankrutuojančioms, siūloma naudoti didelį slenkstį (0.8) mažo disbalanso duomenimis (1 bankrutuojanti ir mažiau kaip 10 veikiančių įmonių) apmokytam modeliui ir mažesnį slenkstį (0.4) didelio disbalanso duomenimis (1 bankrutuojanti ir daugiau kaip 10 veikiančių) apmokytam modeliui.
6. Remiantis atsitiktinių miškų modelių rezultatais, nustatyti reikšmingiausi kintamieji bankroto prognozavimui. Kintamieji susiję su nepriemoka VMI, duomenimis apie darbuotojus, įmonės amžių ir pajamas bei turtą yra svarbiausi, prognozuojant įmonės bankrotą.

Literatūros sąrašas

1. WEI FENG, Wenjiang Huang, Jinchang Ren. Class Imbalance Ensemble Learning Based on the Margin Theory. *Applied Sciences* [interaktyvus]. 2018, 8(5):815 [žiūrėta 2020-02-10]. Prieiga per: doi: http://dx.doi.org/10.3390/app8050815
2. ALBERTO FERNANDEZ, Salvador Garcia ir kt. Learning from Imbalanced Data Sets [interaktyvus]. 2018 [žiūrėta 2020-02-10], ISBN 978-3-319-98073-7. Prieiga per: doi: http://dx.doi.org/10.1007/978-3-319-98074-4
3. ATTENBERG JOSH, Seyda Ertekin. Class Imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications* [interaktyvus]. 2013, 101-149 [žiūrėta 2020-02-10]. Prieiga per: doi: http://dx.doi.org/10.1002/9781118646106.ch6
4. YANG ZHAO, Zoie Shui-Yee Wong ir kt. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection [interaktyvus]. 2018 [žiūrėta 2020-02-10]. Prieiga per: doi: https://doi.org/10.1155/2018/6275435
5. ELENA FEDOROVA, Evgeni V. Gilenko ir kt. Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert. Syst. Appl* [interaktyvus]. 2013, 40, 7285-7293 [žiūrėta 2020-02-10]. Prieiga per: doi: https://doi.org/10.1016/j.eswa.2013.07.032
6. YONG ZHANG, Panpan Fu ir kt. Imbalanced data classification based on scaling kernel-based support vector machine. *Neural Computing and Applications* [interaktyvus]. 2014, 25(3-4), p. 927-935 [žiūrėta 2020-02-10]. Prieiga per: doi: https://doi.org/10.1007/s00521-014-1584-2
7. DAVID VEGANZONES, Eric Severin. An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems* [interaktyvus]. 2018, 112 [žiūrėta 2020-02-10]. Prieiga per: doi: https://doi.org/10.1016/j.dss.2018.06.011
8. SHAONAN TIAN, Yan Yu, Hui Guo. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance* [interaktyvus]. 2015, 52, p. 89-100 [žiūrėta 2020-02-12]. Prieiga per: doi: https://doi.org/10.1016/j.jbankfin.2014.12.003
9. CHIH-FONG TSAI. Feature selection in bankruptcy prediction. *Knowledge-Based Systems* [interaktyvus]. 2009, 22(2), p. 120-127 [žiūrėta 2020-02-12]. Prieiga per: doi: https://doi.org/10.1016/j.knosys.2008.08.002
10. DU JARDIN, PHILIPPE. Bankruptcy prediction models: How to choose the most relevant variables? *Bankers, Markets & Investors* [interaktyvus]. 2009, 98, p. 39-46 [žiūrėta 2020-02-12]. Prieiga per: https://mpra.ub.uni-muenchen.de/44380/
11. MING XU, Chu Zhang. Bankruptcy prediction: the case of Japanese listed companies. Review of Accounting Studies [interaktyvus]. 2009, 14, p. 534-558 [žiūrėta 2020-02-12]. Prieiga per: doi: https://doi.org/10.1007/s11142-008-9080-5
12. KRAL PAVOL, Fleischer M., Stachova M., Nedelova G. Corporate financial distress prediction of slovak companies: z-score models vs. Alternatives [interaktyvus]. 2016 [žiūrėta 2020-02-16]. Prieiga per: http://www.amse.umb.sk/proceedings/KralFleischerStachovaNedelovaSobisek.pdf
13. LOREDANA CULTRERA, Xavier Bredart. Bankruptcy prediction: The case of Belgian SMEs. *Review of Accounting and Finance* [interaktyvus]. 2016, 15(1), p. 101-119 [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.1108/RAF-06-2014-0059

14. NICOLETA BARBUTA-MISU, Elena-Silvia Codreanu. Analysis and Prediction of the Bankruptcy Risk in Romanian Building Sector Companies. *Ekonomika* [interaktyvus]. 2014, 93, p. 131-146 [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.15388/Ekon.2014.2.3542
15. MASSIMILIANO CELLI. Can Z-Score Model Predict Listed Companies' Failures in Italy? An Empirical Test. *An Empirical Test. International Journal of Business and Management* [interaktyvus]. 2015 [žiūrėta 2020-02-16]. Prieiga per: https://researchgate.net. Doi: 10.10.5539/ijbm.v10n3p57
16. BHANDARI SHYAM, Iyer Rajesh. Predicting business failure using cash flow statement based measures. *Managerial Finance* [interaktyvus]. 2013, 39, p. 667-676 [žiūrėta 2020-02-16]. Prieiga per: https://researchgate.net. Doi: 39.667-676.10.1108/03074351311323455
17. OOGHE HUBERT, Sofie De Prijcker. Failure processes and causes of company bankruptcy: A typology. *Management Decision* [interaktyvus]. 2006, 46 [žiūrėta 2020-02-16]. Prieiga per: https://researchgate.net. Doi: 10.1108/00251740810854131
18. GEDIMINAS ŠLEFENDORFAS. Bankruptcy prediction model for private limited companies of Lithuania. *Ekonomika* [interaktyvus]. 2016, 95(134) [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.15388/Ekon.2016.1.9910
19. DAVID ALAMINOS, Agustin Castillo, Manuel Fernandez-Gamez. A Global Model for Bankruptcy Prediction. *PLOS ONE. 11. e0166693* [interaktyvus]. 2016 [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.1371/journal.pone.0166693
20. BAGHER ASGARNEZHAD, Milad Soltani. Designing a bankruptcy prediction model based on account, market and macroeconomic variables (Case Study: Cyprus Stock Exchange) [interaktyvus] 2016, 9 [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.22059/IJMS.2016.55038
21. IVANA PODHORSKA, Mária Mišanková, Katarína Valášková. Searching for Key Factors in Enterprise Bankrupt Prediction: A Case Study in Slovak Republic. *Economics and Culture* [interaktyvus] 2018, 15(1), p. 78-87 [žiūrėta 2020-02-16]. Prieiga per: doi: https://doi.org/10.2478/jec-2018-0009
22. SHEIKH RABIUL ISLAM, William Eberle, Sheikh K. Ghafoor ir kt. Investigating bankruptcy prediction models in the presence of extreme class imbalance and multiple stages of economy. [interaktyvus] 2019. [žiūrėta 2020-02-29]. Prieiga per: https://arxiv.org/abs/1911.09858
23. HORVÁTHOVÁ, JARMILA ir Mokrišová, Martina. Risk of Bankruptcy, Its Determinants and Models. *Risks* [interaktyvus]. 2018, 6(117) [žiūrėta 2020-03-09]. Prieiga per: https://doi.org/10.3390/risks6040117
24. NAGELLA VENKATARAMANA, S.Md.Azash ir K.Ramakrishnaiah. Financial performance and predicting the risk of bankruptcy: a case of selected cement companies in India. *International journal of public administration and management research (ijpamr)* [interaktyvus] 2012, 1(1), p. 40-56 [žiūrėta 2020-03-09]. Doi: RCMSS/IJPAMR/12004
25. KHALIQ, AHMAD. Identifying Financial Distress Firms: A Case Study of Malaysia's Government Linked Companies (GLC) [interaktyvus] 2014. [žiūrėta 2020-03-09]. Prieiga per: www.researchgate.net
26. LUKASON, OLIVER ir Hoffman, Richard. Firm Bankruptcy Probability and Causes: An Integrated Study. *International Journal of Business and Management*. 2014, 9, p. 80-91 [žiūrėta 2020-03-09]. Prieiga per: https://doi.org/10.5539/ijbm.v9n11p80

27. SHI, YIN ir Li, Xiaoni. An overview of bankruptcy prediction models for corporate firms: A Systematic literature review. *Intangible Capital*. [interaktyvus] 2019, 15(114) [žiūrėta 2020-03-09]. Prieiga per: https://doi.org/10.3926/ic.1354
28. SHAZA M. ABD ELRAHMAN ir Ajith Abraham. A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*. 2013, 1, p. 332-340 [žiūrėta 2020-03-09]. Prieiga per: http://www.softcomputing.net/jnic2.pdf
29. HORDRI NUR, Sophiayati Siti ir kt. Handling Class Imbalance in Credit Card Fraud using Resampling Methods. *International Journal of Advanced Computer Science and Applications* [interaktyvus]. 2018, 9 [žiūrėta 2020-03-09]. Priega per: doi: https://doi.org/10.14569/IJACSA.2018.091155
30. SAHIN YUSUF, Bulkan Serol ir Duman Ekrem. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications* [interaktyvus]. 2013, 40, p. 5916–5923 [žiūrėta 2020-03-09]. Prieiga per: doi: https://doi.org/10.1016/j.eswa.2013.05.021
31. ZAREAPOOR MASOUMEH ir Shamsolmoalis Pourya. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. *Procedia Computer Science* [interaktyvus]. 2015, 48, p. 679-686 [žiūrėta 2020-03-09] Prieiga per: doi: https://doi.org/10.1016/j.procs.2015.04.201
32. QIONG GU, Li Zhu ir Zhihua Cai. Evaluation Measures of the Classification Performance of Imbalanced Data Sets. *Computational Intelligence and Intelligent Systems* [interaktyvus]. 2009, 51, p. 461-471 [žiūrėta 2020-05-11] Prieiga per: doi: http://dx.doi.org/10.1007/978-3-642-04962-0_53
33. JAN BRABEC ir Lukas Machlica. Bad practices in evaluation methodology relevant to class-imbalanced problems [interaktyvus]. 2018 [žiūrėta 2020-05-11] Prieiga per: https://arxiv.org/abs/1812.01388
34. ALAA THARWAT. Classification assessment methods. *Applied Computing and Informatics* [interaktyvus]. 2018 [žiūrėta 2020-05-11] Prieiga per: doi: https://doi.org/10.1016/j.aci.2018.08.003.
35. Jūrienė, V. Įmonių bankroto ir restruktūrizavimo bei fizinių asmenų bankroto procesų 2018 m. sausio–gruodžio mėn. apžvalga. *Audito, apskaitos, turto vertinimo ir nemokumo valdymo tarnyba prie Lietuvos respublikos finansų ministerijos* [interaktyvus]. 2019 [žiūrėta 2020-05-17]. Prieiga per: http://www.avnt.lt/assets/Nemokumas/Duomenys-ir-analiz/2018-mAPZVALGA2019-02-01TP.pdf
36. Sergijenko, D. Transportininkų skolos augo, bet bankrotų horizonte nematyti. *Verslo žinios*. 2017, balandis [žiūrėta 2020-05-17]. Prieiga per: https://www.vz.lt/transportas-logistika/2017/04/15/transportininku-skolos-augo-bet-bankrotu-horizonte-nematyti
37. Manuel Fern´andez-Delgado, Eva Cernadas ir Sen´en Barro. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* [interaktyvus]. 2014, 15, p. 3133-3181 [žiūrėta 2020-05-18]. Prieiga per: http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf
38. Colliau Taylor, Rogers Grace ir kt. MatLab vs. Python vs. R. *Business Faculty Publications* [interaktyvus]. 2017, 51 [žiūrėta 2020-05-18]. Prieiga per: https://scholar.valpo.edu/cba_fac_pub/51

39. Marvin N. Wright ir Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* [interaktyvus]. 2017, 77(1) [žiūrėta 2020-05-18]. Prieiga per: doi: https://10.18637/jss.v077.i01
40. Philipp Probst, Marvin Wright ir Anne-Laure Boulesteix. Hyperparameters and Tuning Strategies for Random Forest. [interaktyvus], 2019 [žiūrėta 2020-05-18]. Priegia per: doi: https://doi.org/10.1002/widm.1301