

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
EKONOMETRIJOS BAKALAURO STUDIJŲ PROGRAMA

Agnė Vitkauskaitė

Aukštojo mokslo pasirinkimą lemiantys veiksniai
Factors Influencing the Choice of Tertiary Education

Bakalauro baigiamasis darbas

Darbo vadovė Doc., Dr., Jurgita Markevičiūtė

Vilnius

2020

Turinys

1	Santrauka	3
2	Summary	4
3	Įvadas	5
4	Temos aprašymas	7
5	Metodologija	9
5.1	Logistinė regresija	9
5.2	Modelio tinkamumą vertinantys kriterijai	10
5.3	Faktorinė analizė	11
5.3.1	Pagrindinių komponentų analizė	11
5.3.2	Daugialypė atitikimo analizė	12
5.4	Kryžminė validacija	13
5.5	Tyrime vartojamos sąvokos	14
6	Duomenys	15
6.1	Analizėje naudojami mokinių duomenys	15
6.2	Kintamieji	15
6.2.1	Kintamųjų transformacijos ir reikšmės	17
7	Analizė	19
7.1	Pradinė duomenų analizė	19
7.1.1	Moksleivių stojimas į aukštąsias mokyklas	19
7.1.2	Moksleivių stojimas į aukštąsias mokyklas pagal PUPP vertinimą	20
7.2	Daugianarės logistinės regresijos modeliai	22
7.2.1	1 modelis	23
7.2.2	2 modelis	23
7.2.3	3 modelis	24
7.2.4	1-3 modelių palyginimas	24
7.3	Daugialypė atitikimo analizė	25
7.3.1	SES kintamasis	26

7.3.2	Studijų pajėgumo kintamasis	27
7.3.3	4 modelis	29
7.3.4	5 modelis	29
7.3.5	4-5 modelių palyginimas	29
7.4	2 modelio ir 5 modelio palyginimas	30
8	Išvados ir rekomendacijos	31
9	Priedai	35
9.1	Dešimtbalė vertinimo skalė pagrindiniame ir viduriniame ugdyme	35
9.2	R programos kodas	35

1 Santrauka

Šio darbo tikslas yra pritaikant statistinius duomenų analizės metodus sudaryti statistinį modelį, kuriuo būtų galima įvertinti skirtingų faktorių įtaką mokinio pasirinkimams studijuoti ar ne aukštojoje mokykloje Lietuvoje, iš karto po bendrojo ugdymo mokyklos baigimo. Darbe naudojami metodai apima daugialypės logistinės regresijos analizę, daugialypę atitikimo analizę, kryžminę validaciją. Remiantis gautais logistinės analizės rezultatais, socialinis-ekonominis statusas neturi didelės įtakos mokinių pasirinkimams tęsti ar ne studijas aukštojoje mokykloje, mergaičių šansai studijuoti aukštojoje mokykloje yra didesni nei berniukų bei etninei mažumai priklausantys moksleiviai, turi didesnius šansus tęsti studijas aukštojoje mokykloje.

Raktiniai žodžiai: Aukštasis išsilavinimas, daugialypė logistinė regresija, daugialypė atitikimo analizė, kryžminė validacija, socialinis-ekonominis statusas.

2 Summary

The main goal of this thesis is to make a statistical model using different statistical methods of data analysis. This statistical model evaluates different factors, that influences pupil decisions to continue studies at Lithuanian tertiary education institutions. The main tools used for the analysis are multivariate logistic regression, multiple correspondence analysis, cross-validation. Multivariate logistic regression analysis reveals the result - social-economic status do not have a significant effect on decision to continue higher education studies. Also, girls, and children from ethnic minority have slightly bigger odds to continue studies in tertiary education institution.

Key word: Tertiary education, multivariate logistic regression, multivariate correspondence analysis, cross-validation, social-economic status.

3 Įvadas

Aukštasis išsilavinimas – išsilavinimas, įgyjamas po bendrojo vidurinio išsilavinimo baigus ne trumpesnes kaip 3 metų studijas aukštojoje mokykloje [1]. Aukštuoju mokslu (toliau - AM) ir visa švietimo veikla ugdomi ne tik ateities profesijai ar darbui reikalingi gebėjimai, bet ir asmeninės žmogaus būdo savybės. Aukštojo išsilavinimo (bakalauro) diplomą turintys asmenys dažniau yra dirbantys, gauna didesnę darbo užmokestį, lyginant tik su vidurinį išsilavinimą įgijusiais žmonėmis. Be viso to, aukštasis išsilavinimas asmeniniu individo lygmeniu siejamas ir su sveikesniu, aktyvesniu gyvenimo būdu, geresniais socialiniais įgūdžiais bei pilietiškumu [15]. Taip pat, aukštojo mokslo pagalba tobulinamas ir visuomenės bendras vystymasis, ekonominis ir socialinis pajėgumas bei piliečių aktyvumas. Visa tai reikalinga valstybės ekonomikos raidai plėtoti, jos tvariam augimui bei konkurencingumui skatinti [2]. Kiekvienas studentas yra ateities žmogiškasis išteklius, kuris sukurs pridėtinę vertę valstybei, įnešdamas didesnę naudą ir pajamas į jos bendrą biudžetą. Be didesnių sumokamų mokesčių sumos, dėl aukštesnį išsilavinimą įgijusių piliečių, patiriama mažiau socialinės, sveikatos apsaugos bei viešosios tvarkos užtikrinimo išlaidų [3]. Nepaisant esamų aukštojo išsilavinimo suteikiamų pranašumų tiek asmeniniu žmogaus, tiek valstybės lygiu, nemaža dalis abiturientų nusprendžia nestoti į aukštąsias mokyklas po bendrojo ugdymo mokyklų baigimo [13]. Tokius sprendimus gali lemti kelios priežastys, pavyzdžiui, nesėkmės mokantis, nežinojimas, kurią specialybę pasirinkti ar net šeimos, iš kurios vaikas yra kilęs užimama socialinė bei ekonominė padėtis visuomenėje, t.y. materialinė padėtis [14].

Aukštasis mokslas turi būti prieinamas visiems įgijusiems vidurinį išsilavinimą vienuodai, nepriklausomai nuo užimamos socialinės ir ekonominės padėties šalyje, tačiau atsiranda teigimų, kad nemaža dalis moksleivių, kurie po mokyklos baigimo nusprendė netęsti studijų AM, yra kilę iš žemesnio socialinio - ekonominio statuso (toliau - SES) šeimų. 2015 m. duomenimis, tarp moksleivių iš mažų pajamų šeimų mokslus aukštojoje mokykloje tęsė perpus (25 procentiniais punktais) mažesnė dalis nei tarp likusiųjų [3].

Šio darbo tikslas yra pritaikant statistinius duomenų analizės metodus sudaryti statistinį modelį, kuriuo būtų galima įvertinti skirtingų faktorių įtaką mokinio pasirinkimams studijuoti ar ne aukštojoje mokykloje Lietuvoje, iš karto po bendrojo ugdymo mokyklos baigimo. Taip pat patvirtinti arba atmesti hipotezę, kad žemo SES moksleiviai yra linkę rečiau rinktis studijas aukštojoje mokykloje.

Darbo išskirtinumas yra tai, jog pritaikant daugialypę atitikimo analizę (angl. – multiple correspondence analysis) sumažinta naudojamų duomenų dimensija. Duomenų dimensijos sumažinimu prarandama dalis informacijos, tačiau tai į modelį leido įtraukti ženkiai mažiau nepriklausomų kintamųjų. Modelyje, kuris sukurtas su ženkiai mažiau nepriklausomų kintamųjų, bendras teisingai klasifikuotų reikšmių procentas nesumažėja, lyginant su modeliu, į kurį įtraukta daugiau nepriklausomų kintamųjų. Analizei atlikti naudoti duomenys yra nauji, todėl gavome patikimus, aktualius, o ne senesnių metų rezultatus.

4 Temos aprašymas

Studijų tęsimo klausimas - dažnas analizės objektas tiek individo, tiek valstybiniu lygiu. Asmens lygiu jis sprendžiamas iš kelių perspektyvų, pavyzdžiui: asmeniniai su studijomis susiję kriterijai, tokie kaip geografinė aukštosios mokyklos padėtis, pomėgiai, studijų kokybė, norai. Lietuvos lygiu studijų klausimas dažnai analizuojamas Vyriausybės strateginis analizės centro – Strata (ankščiau - Mokslo ir studijų stebėsenos ir analizės centras (Mosta)). Centras 2014 m. atliko aukštojo mokslo sistemos suinteresuotų šalių tyrimą - „Studijų pasirinkimo motyvai ir pasirengimas studijoms“ [4]. Tyrime analizuojami skirtingi veiksniai, lemiantys studijų pakopų, aukštosios mokyklos, studijų programos pasirinkimus.

Tyrime teigiama, kad renkantis studijas svarbu žinoti studijų vietą, įvertinti savo gabumus, tinkamai pasiruošti studijoms. Tačiau, nepaisant asmeninių norų, moksleivių teigimu, 42 procentai rinkęsi valstybės finansuojamą vietą. Taip pat, analizėje gauti rezultatai, kad po valstybinių brandos egzaminų, mokama ar nemokama vieta reiškia mažiau. Pagal tyrime atliktą analizę, didžiausią reikšmę pasirinkimams, kur tęsti mokslus, turėjo studijų programa (46 procentai), studijų finansavimo pobūdis liko antroje vietoje (21 procentas). Net keturi penktadaliai tyrime dalyvavusių moksleivių linkę rinktis širdžiai mielą specialybę, o ne specialybę, užtikrinančią materialią gerovę. Šis mąstymo pobūdis pastebimas tarp geresniais pažymiais besimokančių moksleivių. Be to, nemokama studijų vieta svarbesnė mažesniuose miestuose gyvenantiems bei geriau besimokantiems moksleiviams.

Kitas Strata 2018 m. atliktas tyrimas - „Aukštojo mokslo prieinamumas: ar skirtingo SES vaikams sudarome lygias galimybes įgyti išsilavinimą“ [3]. SES - santykinė asmenų padėtis socialinėje sistemoje. Analizėje apibrėžiami trys skirtingi SES šaltiniai: ekonominiai, kultūriniai ir socialiniai. Tyrime gauti rezultatai - valstybinius brandos egzaminus gerai¹ išlaikę abiturientai nepriklausomai nuo šeimos pajamų į aukštąją mokyklą įstojo 8 atvejais iš 10. Tačiau nepaisant to, iš žemas pajamas gaunančių šeimų kilę moksleiviai rečiau įgijo vidurinį išsilavinimą, rečiau laikė bei prasčiau išlaikė valstybinius brandos egzaminus. Baigę 10 klasių, žemo SES šeimų mokiniai daugiau nei dvigubai dažniau rinkosi tęsti mokslus profesinėse mokyklose. Analizėje taip pat gautos išvados, kad beveik kas trečias iš žemų pajamų šeimos kilęs moksleivis Lietuvoje neįgijo nei brandos atestato nei profesinės kvalifikacijos, kadangi neįkvalifikuotų darbuotojų pasiūla smarkiai viršija paklausą, buvo teigiama,

¹Gerau egzaminus išlaikę moksleiviai laikomi tie, kurie iš VBE vidutiniškai surinko daugiau nei 70 balų.

kad moksleivių ateities galimybės yra siauros, jie nesugebės įsitvirtinti darbo rinkoje ir žemas SES bus perduodamas iš kartos į kartą. Be to, tyrime gauti rezultatai rodo, kad žemo SES šeimų jaunuoliai dažniau rinkosi mažesnių balų reikalaujančias aukštojo mokslo įstaigas, bet tuo pat metu, tų studijų programų baigimas disponuos mažesnes pajamas ateityje. Praėjus metams po studijų baigimo, absolventų iš žemų pajamų šeimų pajamos buvo mažesnės 12 procentinių punktų, o aukštos kvalifikacijos reikalaujančiuose darbuose jie taip pat įsidarbindavo 8 procentiniais punktais rečiau.

5 Metodologija

Darbe rezultatai gauti naudojant administracinius duomenis, jiems pritaikyta:

- Aprašomosios statistikos vizualinė analizė.
- Daugialypės logistinės regresijos analizė.
- Faktorinė analizė:
 - Pagrindinių komponentų analizė (PKA) (angl. - Principal Component Analysis (PCA)).
 - Daugialypė atitikimo analizė (DAA) (angl. - Multiple correspondence analysis (MCA)).
- Logistinė regresija naudojant gautas DAA pagrindines komponentes.
- Kryžminė validacija (angl. - cross-validation arba out-of-sample testing).

Tolesniuose skyriaus poskyriuose detalizuoti ir plačiau aprašyti darbe naudojami analizės metodai.

5.1 Logistinė regresija

Binarinė logistinė regresija yra statistinės analizės metodas, kai modeliuojamas nepriklausomas - dvireikšmis kintamasis Y . Dvireikšmis kintamasis gali įgyti dvi reikšmes, dažniausiai - 0 ir 1. Modelis yra aprašomas tikimybe:

$$P(Y = 1) = \frac{e^{z(x)}}{1 + e^{z(x)}},$$

čia $z(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$,

$x_1, x_2, \dots, x_k$ - kovariantės (kiekybinės arba kokybinės),

k - požymių skaičius duomenyse.

Modelio tinkamumas vertinamas remiantis rodikliais: *tikėtinumo santykio kriterijus*, *Voldo kriterijus regresoriams*, *Akaike informacinis kriterijus (AIC)*, *determinacijos (pseudo) koeficientas* bei *klasifikacinė lentelė* (rodikliai apibrėžti 5.2 skyriuje) [12].

Vertinant logistinės regresijos parametrus žiūrima į koeficiento reikšmę prie regresoriaus. Jei koeficientas prie regresoriaus teigiamas, tai jam didėjant, tikimybė $P(Y = 1)$

didėja. O jei koeficientas neigiamas, tai regresoriui didėjant, tikimybė $P(Y = 1)$ mažėja (tikimybė $P(Y = 0)$ didėja).

Modelio pagalba galime įvertinti kintamųjų svarbą. Koeficiento β_k eksponentė (e^{β_k}) vadinama *galimybių santykiu* (angl. *odds ratio*). Apskaičiavus galimybių santykius galime matyti, kiek pasikeis tikimybė $P(Y = 1)$, kai x_k reikšmė padidės vienetu, o kitos kovariantės liks nepakitusios [6] [7].

Atliekant daugialypę logistinę regresiją, modeliuojamas priklausomas kintamasis turi tenkinti apribojimus. Pavyzdžiui, duomenyse negali vyrauti viena iš kintamojo įgyjamų reikšmių, geriausiai kai vienetas/nulis sudaro ne mažiau nei penktadalį reikšmių [12].

5.2 Modelio tinkamumą vertinantys kriterijai

- *Tikėtinumo santykio kriterijus* - χ^2 . Jei gauti logistinės regresijos modelio koeficientai $\hat{\beta}_k$ yra nuliai, tai modelis prognozavimui yra netinkamas [18]. Kriterijus parodo, ar bent vienas į logistinę regresiją įtrauktas nepriklausomas kintamasis yra statistiškai reikšmingas [12]. Statistinė hipotezė:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1 : \text{bent vienas } \beta_k \neq 0 \end{cases}$$

jei nėra pakankamų įrodymų amesti H_0 , tai logistinės regresijos modelis nėra tinkamas duomenims. Jei H_0 atmetame, tai modelis, į kurį įtraukti regresoriai yra geresnis, nei modelis be regresorių [18].

- *Voldo kriterijus regresoriams*. Voldo kriteriju nurodo ar konkretaus koeficiento reikšmė lygi nuliui ($\hat{\beta}_k = 0$). Statistinė hipotezė:

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}$$

Jei H_0 neatmetama, tai regresorius yra statistiškai nereikšmingas. Tokiais atvejais nepriklausomas kintamasis dažniausiai yra šalinamas iš modelio lygties [12] [18].

- *Informacinis indeksas*. Analizėna naudotas *Akaike informacinis kriterijus (AIC)*. AIC remiasi didžiausio tikėtinumo įvertinių asimptotine teorija [19]. Jis naudojamas mo-

delių tarpusavio palyginimui, kuo mažesnė kriterijaus reikšmė, tuo modelis laikomas geresniu [12].

- *Determinacijos (pseudo) koeficientas*. Tiesinėje regresijoje norint įvertinti regresorių ir regresanto prilausomybės dydį yra skaičiuojamas determinacijos koeficientas R^2 . Logistinėje regresijoje skaičiuojami įvairūs koeficiento analogai. Šiame darbe naudojamas *Makfadeno (angl.- McFadden)* pseudodeterminacijos koeficientas R^2 . Koeficientas įgyja reikšmės iš intervalo $[0,1]$. Kuo reikšmė didesnė, tuo modelis labiau tinka duomenims [12] [18].
- *Klasifikacinė lentelė*. Ja tikrinama, kokia dalis priklausomo kintamojo yra teisingai klasifikuota. Stebėjimams yra prognozuojama Y reikšmė ir žiūrima ar reikšmė sutapo su tikrąja. Kuo didesnis procentas sutampančių reikšmių, tuo modelis yra geresnis [12].

Prognozavimui taikoma taisyklė:

Į formulę $z(\vec{x}) = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ įstatomos $x_1, \dots, x_k$ reikšmės. Jei $z(\vec{x}) > 0$, tai prognozuojama, kad $Y = 1$. Jei $z(\vec{x}) < 0$, tai prognozuojama, kad $Y = 0$. Duomenų reikšmės ir prognozės pateikiamos klasifikacine lentele. Jeigu prognozės nėra tikslios, tai logistinė regresija nėra taikytina [18].

5.3 Faktorinė analizė

5.3.1 Pagrindinių komponentų analizė

Pagrindinių komponentų analizė (PKA) - statistinis metodas, naudojama tada kai susiduriama su dideliu koreliuotų duomenų rinkiniu. Analizė leidžia apibendrinti turimus kintamuosius ir naudojant tiesines transformacijas gauti mažesnės dimensijos duomenis. Gauti duomenys yra pakankamai reprezentatyvūs, jie atspindi didžiausią galimą analizėje naudojamų duomenų variaciją [5].

Pagrindinė PKA idėja - naudojant tiesinę transformaciją sumažinti duomenų matmenų skaičių, vėliau atsisakant po transformacijos gautų, mažiausią dispersiją turinčių, komponentų.

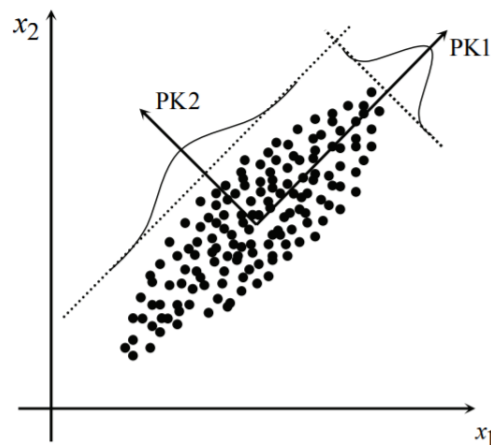
Komponentė einanti per centrinį duomenų tašką ir turinti didžiausią dispersiją yra vadinama *pirmąja pagrindine komponente* (PK1). Centrinis taškas yra tas, kurio kompo-

mentės yra duomenų aibę sudarančių taškų atskirų komponentių vidurkiai. Visų analizėje žiūrimų taškų atstumas iki pirmos pagrindinės komponentės yra minimalus.

Antroji pagrindinė komponentė (PK2) taip pat eina per centrinį tašką ir yra ortogonaliai pirmajai pagrindinei komponentei.

Pavyzdys, kaip gali atrodyti pirmos dvi pagrindinės komponentės, pavaizduotas 1 pav. [8]. Analizės metodu sukurtos komponentės tarpusavyje nekoreliuoja [18].

1 pav.: Pirmoji (PK1) ir antroji (PK2) pagrindinės komponentės



5.3.2 Daugialypė atitikimo analizė

Daugialypė atitikimo analizė (DAA) taip pat kaip ir PKA, faktorinės analizės metodas, tačiau pritaikytas kokybiniais kintamiesiems. Čia taip pat stebima kintamųjų priklausomybė, tačiau priešingai nei PKA, ji negali būti matuojama koreliacijos koeficientu. Dažnai siekiant išsiaiškinti priklausomybes, asociacijos tarp kintamųjų yra vizualizuojamos.

DAA gali būti vertinama kaip PKA tada, kai skaičiavimams naudojama pilna disjunktyvinė lentelė (angl. - complete disjunctive table (CDT)). Lentelės eilutės žymi asmenis (žym.- i), o stulpeliai - duomenų kategorijas (žym.- k). Gauta lentelė sudaryta iš y_{ik} reikšmių, kur $y_{ik} = 1$, jei asmeniui i priskiriama k kategorija ir $y_{ik} = 0$, jei kategorija k asmeniui i nepriskiriama². Apibrėžkime:

- J - kintamųjų skaičius

²Lentelė nėra tiesiogiai sukurama analizės metu programinės įrangos pagalba, tačiau DAA yra paremta jos skaičiavimu

- K_j - kategorijų skaičius kintamojo j
- K - bandras kategorijų skaičius ($K = \sum_j K_j$)

Pilna disjunktyvinė lentelė pavaizduota 2 pav. [20].

2 pav.: Pilna disjunktyvinė lentelė

	1		j			J		Σ
		1	k	K_j				
1								
i			y_{ik}					
l			0 1 0 0 0					
I								
							J	

Tada, tegul p_k žymi asmenų, turinčių k kategoriją dalį. Transformuota pilna disjunktyvinė lentelė:

$$x_{ik} = \frac{y_{ik}}{p_k - 1}.$$

Kadangi PKA stulpeliai turi būti centruoti, tai iš p_k atimamas 1. Jei transformuotai pilnai disjunktyvinei lentelei pritaikysime nestandartizuotą PKA, gausime DAA rezultatus [20].

5.4 Kryžminė validacija

Statistikos metodas, kurio esmė yra susikurti duomenų imtį, padedančią įvertinti modelio prognozavimo galimybes. Įgyvendinimo algoritmas - suskirstyti visą duomenų masyvą į dvi atsitiktinai pasirinktas dalis. Viena duomenų dalis naudojama modelio prognozei įvertinti, o kita jo apmokymui [16]. Metodo pagalba galime įvertinti kaip sudarytas modelis gali prognozuoti reikšmes. Vertinant prognozės tikslumą naudojami duomenis, kurie nebuvo įtraukti į modelio kūrimui naudojamų duomenų imtį [17].

5.5 Tyrime vartojamos sąvokos

- *Socioekonominis statusas (SES)* - santykinė padėtis socialinėje sistemoje, apibrėžiama pagal galimybes naudoti ar kontroliuoti visuomenėje vertinamus išteklius, tokius kaip turtas, galia ar statusas [3]. Kintamasis gali būti paaiškinamas asmeninėmis ar namų ūkio pajamomis, išsilavinimu, darbe užimama kvalifikacija ar profesija [9].
- *Aukštoji mokykla (AM)* - studijų ir mokslo įstaiga, rengianti aukštąjį išsilavinimą turinčius piliečius [10]. Tyrime tai gali būti universitetas, kolegija, aukštoji mokykla ar akademija.
- *Valstybinis brandos egzaminas (VBE)*.
- *Pagrindinio ugdymo pasiekimų patikrinimas (PUPP)*.
- *Valstybės nefinansuojama vieta (VNF)*.
- *Valstybės finansuojama vieta (VF)*.

6 Duomenys

Tyrimui atlikti naudoti administraciniai duomenys, gauti iš Vyriausybės strateginio analizės centro (Strata). Informacinių technologijų centro suteiktu leidimu, Strata turi teisę naudotis duomenų rinkiniu. Pirminis duomenų šaltinis - Švietimo valdymo informacinė sistema (ŠVIS), savo ruožtu, duomenis surenka iš daugybės kitų registrų ir valdo juos savininkų teisėmis.

Analizėje naudojami duomenys apima 2010-2011 - 2019-2020 mokslo metus. Duomenyse atsispindi informacija apie moksleivių pasiekimus mokyklose, laikytų PUPP egzaminų rezultatus, tolimesnius studijų pasirinkimo rodiklius, tėvų socialinį - ekonominį statusą bei kita. Atliekant tyrimą duomenims taikomi atitinkami apribojimai aprašomi toliau esančiuose skyriuose ir jų poskyriuose.

6.1 Analizėje naudojami mokinių duomenys

Į visą duomenų imtį yra įtraukti profesinėse ugdymo (PU) ir bendrojo ugdymo (BU) mokymosi įstaigose studijuojantys moksleiviai. Tik 4 procentai 10-oje klasėje PU mokyklose besimokančių mokinių iš karto po mokyklos baigimo stojo į aukštąją mokyklą Lietuvoje. PU įstaigų mokiniai sudaro kiek mažiau nei 19 procentų visų imtyje esančių 10-os klasės moksleivių, jie iš tyrimo naudojamos mokinių aibės duomenų bus pašalinami. Tyrime koncentruojamasi į BU mokyklų studentus.

Iš tyrimo naudojamų duomenų atmesime ir *etninė_mažuma* kintamojo *NA* reikšmes (kintamasis apibrėžtas 6.2.1 skyriuje). Pradiniuose duomenyse tai sudarė tik 0.09 procento visos imties.

6.2 Kintamieji

Analizėje naudojami kintamieji pateikti 1 lentelėje. Visų tyrimo kuriamų daugialypės logistinės regresijos modelių priklausomas kintamasis *Y* (*įstojo*) yra moksleivio apsisprendimas studijuoti AM iš karto po bendrojo ugdymo mokyklos baigimo be pertraukos Lietuvoje. Jis gali įgyti dvi reikšmes, 1 - kai mokinsys tęsia studijas Lietuvos aukštojoje mokykloje bei 0 - priešingu atveju, kai studijų Lietuvos aukštojoje mokykloje mokinsys netęsia.

Atliekant daugialypę atitikimo analizę sukurtos komponentės ir į jas įtraukiami kintamieji aprašomi 7.3.1 bei 7.3.2 skyriuose.

1 lentelė: Kintamieji

Kintamasis	Apibūdinimas	Tipas	Galimos reikšmės
Istojo	Indikatorius, kad moksleivis įstojo į AM be pertraukos po BU mokyklos baigimo Lietuvoje	Dvireikšmis	0 arba 1
Pajamos	Suminių tėvų pajamų kvartiliai	Kategorinis	
Lytis	Mokinio lytis	Dvireikšmis	mergaitė/ berniukas
Etninė_mažuma	Vaiko priklausomybė etninei mažumai	Dvireikšmis	taip/ne
Gyvenvietė	Gyvenamoji vieta	Dvireikšmis	miestas/ kaimas
Bent_vienas_aukšt_kval	Indikatorius, kad bent vienas iš tėvų dibi aukštos kvalifikacijos darbą	Dvireikšmis	0 arba 1
Kvalif_tėtis	Tėčio užimama kvalifikacija darbe	Kategorinis	
Kvalif_mama	Mamos užimama kvalifikacija darbe	Kategorinis	
Mat_PUPP_vid	Matematikos PUPP rezultatas, kai NA reikšmės pakeistos vidurkiu	Kiekybinis	[1,10]
Liet_PUPP_vid	Lietuvių kalbos PUPP rezultatas, kai NA reikšmės pakeistos vidurkiu	Kiekybinis	[1,10]
Mat_PUPP_0	Matematikos PUPP rezultatas, kai NA reikšmės pakeistos 0	Kiekybinis	[0,10]
Liet_PUPP_0	Lietuvių kalbos PUPP rezultatas, kai NA reikšmės pakeistos 0	Kiekybinis	[0,10]
Mat_PUPP_kat	Matematikos PUPP rezultatai kategorijomis	Kategorinis	
Liet_PUPP_kat	Lietuvių kalbos PUPP rezultatai kategorijomis	Kategorinis	
Mat_PUPP_laikė	Indikatorius, nurodantis ar mokinys laikė matematikos PUPP egzaminą	Dvireikšmis	0 arba 1
Liet_PUPP_laikė	Indikatorius, nurodantis ar mokinys laikė lietuvių kalbos PUPP egzaminą	Dvireikšmis	0 arba 1
Egz_suma	Laikytų egzaminų skaičius	Kiekybinis	[0,23]

6.2.1 Kintamųjų transformacijos ir reikšmės

Norint tinkamai atlikti ir įvertinti logistinės regresijos modelius, duomenims buvo pritaikytos tam tikros transformacijos:

- Matematikos ir lietuvių kalbos PUPP praleistos *NA* reikšmės keičiamos trimis skirtingais būdais, tam kad būtų galima sukurti kuo patikimesnį ir duomenims labiau pritaikytą logistinės regresijos modelį:
 1. Matematikos ir lietuvių kalbos PUPP *NA* reikšmės keičiamos visų, atitinkamai matematikos arba lietuvių kalbos PUPP, reikšmių vidurkiais. Gauti kintamieji: *Mat_PUPP_vid*, *Liet_PUPP_vid* (1 lentelė).
 2. Matematikos ir lietuvių kalbos PUPP reikšmės keičiamos į kategorinį kintamąjį. Tai daroma pagal vertinimo skalės bendrajame ugdyme skirstymą. Taip sukuriamos kategorijos ir kartu pakeičiamos *NA reikšmės*. *NA* priskiriama egzaminą nelaikiusiųjų kategorijai. Kitos kategorijos apibrėžtos prieduose esančioje 8 lentelėje. Gauti kintamieji: *Mat_PUPP_kat*, *Liet_PUPP_kat* (1 lentelė).
 3. Matematikos ir lietuvių PUPP *NA* reikšmės keičiamos 0. Gauti kintamieji: *Mat_PUPP_0*, *Liet_PUPP_0* (1 lentelė).
- *Etninė_mažuma*. Kintamasis perkoduotas į dvireikšmį. Kai kintamasis įgyja reikšmę „taip“, tai mokinys priklauso etninei mažumai, jo ugdymo kalba nebuvo lietuvių 10-oje klasėje. Kai kintamasis įgyja reikšmę „ne“ - mokinys nepriklauso etninei mažumai.
- *Gyvenvietė*. Kintamasis perkoduotas į dvireikšmį, jis įgyja reikšmes: „miestas“ arba „kaimas“. Kintamasis įgyja reikšmę „kaimas“ kai asmuo mokėsi kaime 10-oje klasėje ir reikšmę „miestas“, kai mokinys mokėsi mieste, miestelyje arba duomenyse apibrėžtoje gyvenvietėje - geležinkelio stotis.
- *Liet_PUPP_laikė*, *Mat_PUPP_laikė*. Dvireikšmiai kintamieji, 1 - kai mokinys laikė atitinkamai lietuvių kalbos arba matematikos egzaminą ir 0 - kai laikyto egzamino kintamojo reikšmė buvo *NA*.
- *Bent_vienas_aukšt_kval*. Dvireikšmis kintamasis, kuris lygus 1, kai mamos arba tėčio darbo kvalifikacija yra aukšta ir 0, kai nei vienas iš tėvų nedarbo aukštos kvalifikacijos darbu.

- *Kvalif_tėtis, kvalif_mama*. Kategoriniai kintamieji sukurti iš darbo profesijos kodų remiantis Lietuvos profesijų klasifikatoriumi [11]. Čia kintamasis gali įgyti reikšmes:
 - *Neturi*. Kai atitinkamai vaikas neturi mamos arba tėčio.
 - *Nedirbo*. Kai atitinkamai mama arba tėtis neturėjo darbo.
 - *Nepriskirtos kvalifikacijos*. Kai atitinkamai mama arba tėtis dirbo nepriskirtos kvalifikacijos darbą.
 - *Nekvalifikuotas*. Kai atitinkamai mama arba tėtis dirbo nekvalifikuotą darbą.
 - *Žemos kvalifikacijos*. Kai atitinkamai mama arba tėtis dirbo žemos kvalifikacijos darbą.
 - *Aukštos kvalifikacijos*. Kai atitinkamai mama arba tėtis dirbo aukštos kvalifikacijos reikalaujantį darbą.
- *Pajamos*. Pajamų kintamasis gautas sudėjus mamos ir tėčio sumines metines pajamas iš samdomojo darbo (neatskaičius mokesčių) metais, kai vaikas buvo 10-oje klasėje. Tyrime naudosime šį kintamąjį suskaidytą į kvartilius, imant jį ne kaip skaitinį, o kaip kategorinį kintamąjį. Kategorinis kintamasis įgyja 5 skirtingas reikšmes. Gautų kvartilių reikšmės pateikiamos 2 lentelėje.

2 lentelė: Pajamų kintamojo kvartiliai

Kvartilis	Pajamos eurais
Nedirbo samdomojo darbo	0
Q1	4352.7
Q2	8620.1
Q3	15015.7
Q4	807868.1

7 Analizė

Skyrius skaidomas į poskyrius. Pirmajame atlikta pirminė, vizualinė duomenų analizė, peržvelgiamos stojimo į aukštąsias mokyklas tendencijos. Antrajame kuriami trys duomenims pritaikyti logistinės regresijos modeliai, žiūrima į jų tinkamumą duomenims. Trečiajame poskyryje naudojant DAA, sukurti SES bei mokymosi pajėgumą atspindintys kintamieji, vėliau jiems pritaikytas daugialypės logistinės regresijos modeliavimas.

7.1 Pradinė duomenų analizė

7.1.1 Moksleivių stojimas į aukštąsias mokyklas

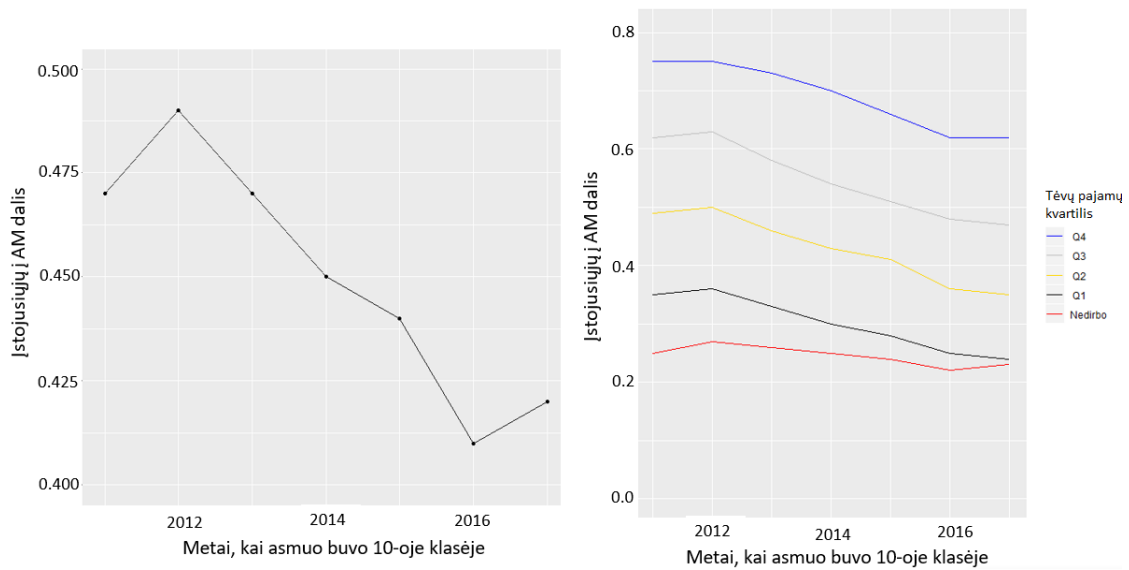
Tyrimo pagrindinis klausimas yra susijęs su moksleivių pasirinkimu tęsti ar ne studijas aukštojoje mokykloje Lietuvoje, iš karto po bendrojo ugdymo mokyklos baigimo. Stebimas 2011-2017 metų stojimas į AM. Procentinė stojančiųjų dalis 2012-2016 metais sumažėjo aštuoniais procentiniais punktais (3 pav. kairėje esantis grafikas).

Kitas tyrimo klausimas: išsiaiškinti SES įtaką moksleivių pasirinkimui studijuoti aukštojoje mokykloje. Norint tai ištirti svarbu pažiūrėti ir į stojančiųjų procentines dalis atsižvelgiant į SES kintamąjį. Čia SES kintamasis - *pajamos* (plačiau apie kintamąjį rašoma 6.2 skyriuje).

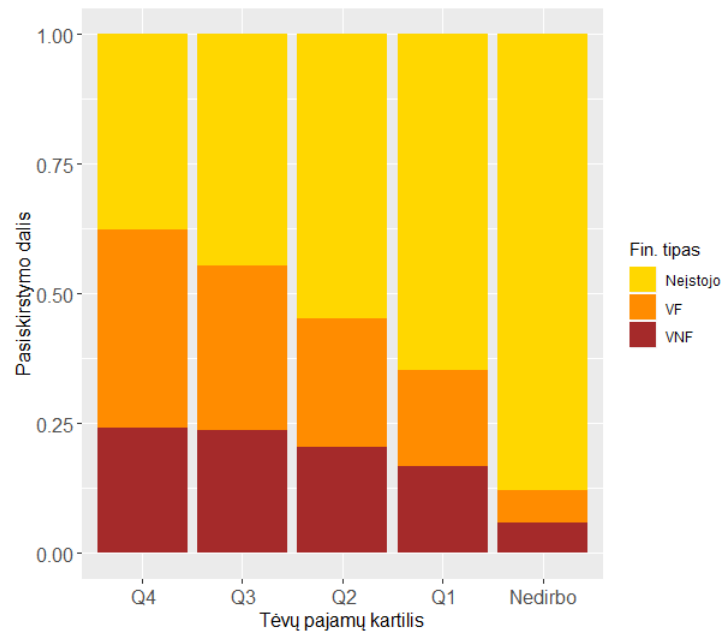
3 pav. dešinėje pateiktame grafike pavaizduota stojančiųjų moksleivių į AM dalis pagal skirtingą šeimų socialinį-ekonominį statusą. Nedarbančių tėvų vaikai vidutiniškai 44,4 procentiniais punktais rečiau įstoja į AM, didžiausias atotrūkis stebimas 2011 metais, jis siekia net 50 procentų. Tačiau kiekvienais metais atotrūkis darosi vis mažesnis. Tėvų pajamoms mažėjant, procentinė stojančiųjų į AM dalis taip pat mažėja. Be to, įstojusiųjų į AM dalies atotrūkis tarp mažiausias pajamas gaunančių tėvų vaikų ir nedarbančių tėvų vaikų su metais vis mažėja, ties 2017 metais jis tampa labai minimalus - tik 1 %.

Stebint mokinių pasiskirstymą tarp skirtingas pajamas uždirbančių tėvų ir įstojimo į aukštąsias mokyklas finansavimo pobūdžio matome, kad mažesnės tėvų pajamos lemia ir mažesnę įstojimą į AM procentą. Taip pat palaipsniui pajamoms mažėjant, mažėja ir į valstybės finansuojamą vietą įstojusių abiturientų procentinė dalis (4 pav.).

3 pav.: Mokinių stojimas į AM



4 pav.: Mokinių stojimas į AM pagal finansavimo pobūdį



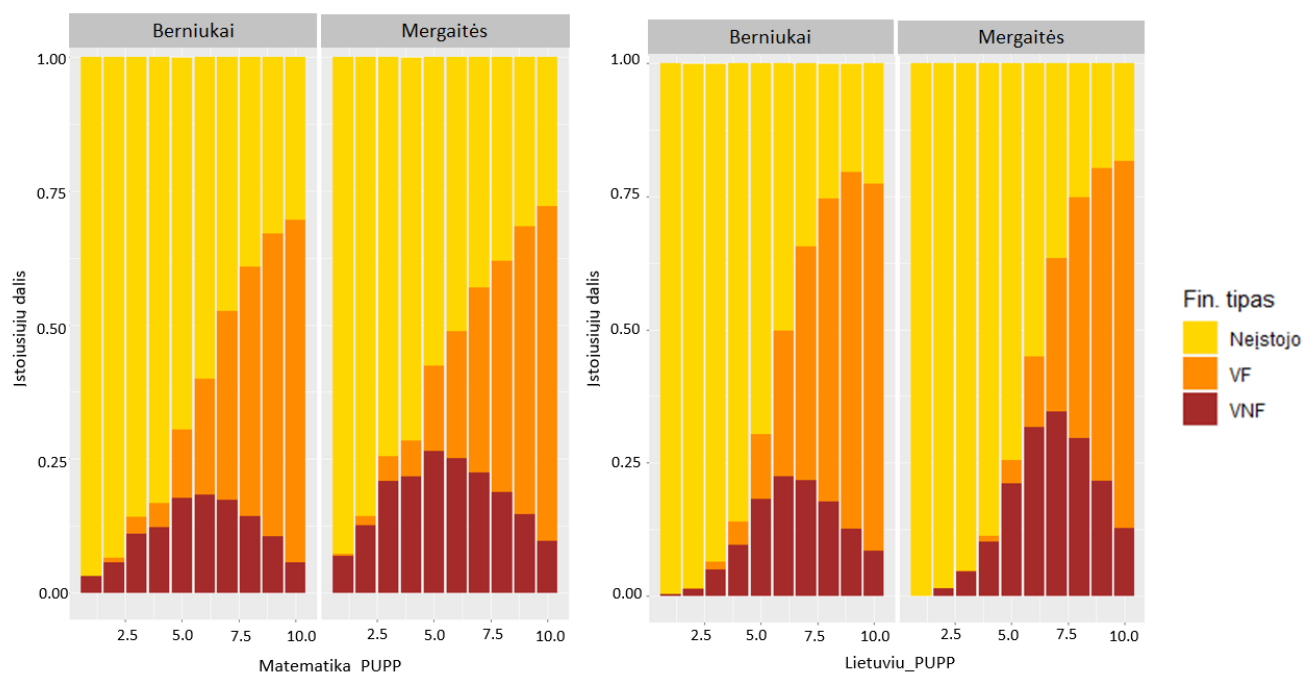
7.1.2 Moksleivių stojimas į aukštąsias mokyklas pagal PUPP vertinimą

Svarbus veiksnys, lemiantis abiturientų stojimą į aukštąsias mokyklas iš karto po bendrojo ugdymo mokyklos baigimo, turėtų būti moksleivio studijų rezultatai pasiekti bendrojo ugdymo mokykloje. Kadangi tyrime stebime 10 klasės moksleivius, 5 pav. pateikiama įstojusiujų į AM moksleivių dalis, aproksimuota matematikos bei lietuvių kalbos PUPP egzaminų rezultatais. Grafike stebima, kad didesnė dalis mergaičių (vidutiniškai 5.645 pro-

centiniais punktais), nei berniukų įstojo į valstybės nefinansuojamą vietą (VNF) Lietuvos aukštojoje mokykloje. Tačiau, vidutiniškai 2.54 procentiniais punktais berniukai dažniau yra pakviečiami studijuoti valstybės finansuojamoje vietoje. Iš grafikų atrodo, kad gan panašus bendras stojančiųjų procentas yra tarp abiejų lyčių.

Remiantis vizualine analize (5 pav.), stebima tendencija: kuo didesnis matematikos ir lietuvių kalbos PUPP pažymys, tuo labiau abiturientas yra linkęs tęsti studijas aukštojoje. Taip pat, didesnis procentas aukščiausius PUPP įvertinimus gaunančių moksleivių studijuoja valstybės finansuojamoje vietoje, lyginant su vidutinius pažymius gaunančiais studentais.

5 pav.: Mokinių stojimas į AM pagal PUPP pasiekimus



7.2 Daugianarės logistinės regresijos modeliai

Sąlyga, kad vienos iš reikšmių (vieneto/nulio) dominavimas būtų ne mažesnis nei 20% išpildyta, kadangi vienetai sudaro 34 %, o nuliai 66 % priklausomo kintamojo reikšmių.

Visuose daugianarės logistinės regresijos modeliuose vertinamas priklausomas kintamasis Y apibrėžiamas 6.2 skyriuje.

Norint įvertinti mokinių pasirinkimo tęsti studijas aukštojoje mokykloje priklausomybę nuo bendrojo ugdymo mokykloje pasiektų rezultatų, į logistines regresijas įtraukiami matematikos ir lietuvių kalbos PUPP rezultatai. Atitinkamai matematikos ir lietuvių kalbos pasiekimų įvertinimų rezultatuose, 37 ir 52 procentai visų reikšmių sudarė praleisti stebėjimai NA . Praleistos reikšmės atsirado dėl to, kad mokinys nelaikė PUPP egzamino. Siekiant pasiekimus tinkamai įvertinti, praleistoms reikšmėms pritaikytos skirtingų tipų transformacijos (praleistų reikšmių NA keitimo būdai detalizuoti 6.2.1 skyriuje). Pritaikant skirtingus praleistų reikšmių keitimo būdus, pirmiausia sukurti trys daugialypės logistinės regresijos modeliai. Vėliau, remiantis logistinės regresijos modelio parametrais, įvertinamas jų tinkamumas duomenims.

Norint įvertinti studijų pasiekimų įtaką logistinėje regresijoje, naudota tik matematikos ir lietuvių kalbos PUPP rezultatai. Papildomi moksleivio studijų pajėgumą atspindintys kintamieji, tokie kaip valstybinių brandos egzaminų rezultatai, manomai, stipriai koreliuoja su PUPP rezultatais, todėl jų įtraukimas į modelius galimai lemtų nepatikimus įvertinių ženklus logistinėje regresijoje. Žiūrėti PUPP rezultatus pasirinkta todėl, kad tyrime stebima 10-oje klasėje buvę moksleiviai.

Siekiant įvertinti šeimos SES įtaką mokinių sprenimui rinktis ar ne studijas aukštojoje mokykloje į regresijas įtraukti kintamieji: *pajamos*, *bent_vienas_aukšt_kval*, *kvalif_mama*, *kvalif_tėtis* ir *gyvenvietė*. Kintamieji pateikti 1 lentelėje.

Dėl duomenų kokybės atmetus kelis netinkamus stebėjimus, logistiniai modeliai sudaromi naudojant 2013-2014 - 2019-2020 mokslo metų duomenis. Visų modelių apmokymo imtis sudaro 319607 stebėjimai. Pritaikant kryžminės validacijos metodą, 4000 (1.24%) atsitiktinai parinktų stebėjimų išskirta modelių testavimui.

7.2.1 1 modelis

Modelio parametrai pateikti 4 lentelėje.

Pirmasis logistinės regresijos modelis sukurtas įtraukiant matematikos ir lietuvių kalbos PUPP NA reikšmes pakeistas rezultatų vidurkiais (kintamieji plačiau aprašomi 6.2 skyriuje ir jo poskyryje).

Modelis tiko duomenims, tikėtinumo santykio kriterijus $\chi^2 = 207714$, p-reikšmė < 0.001 . *Liet_PUPP_vid* buvo statistiškai nereikšmingas kintamasis (Voldo kriterijaus p-reikšmė didesnė nei $\alpha = 0.05$), tačiau jis iš regresijos nešalinamas, kad galėtume tinkamai palyginti modelius. Taip pat mamos kvalifikacijos skirtingas kategorijas atspindintys kintamieji: *kvalif_mama_nekval* ir *kvalif_mama_aukštos_kval* statistiškai nereikšmingi (Voldo kriterijaus p-reikšmės didesnės nei $\alpha = 0.05$). Kadangi kintamieji yra dvi atskiros *kvalif_mama* kintamojo kategorijos, jų taip pat nenaikinome. Determinacijos pseudokintamojo reikšmė didelė - $R^2 = 0.507$.

Klasifikacinė lentelė (3 lentelė) buvo sudaryta naudojant 4000 atsitiktinai paimtų stebinių. Taikant modelį atsitiktinai paimtai imčiai gauta, kad teisingai klasifikuoti 81.6 procentai į AM stojusių ir 89 procentai nestojusiųjų studentų. Bendras teisingų klasifikavimų procentas: 86.525 (4 lentelė).

7.2.2 2 modelis

Modelio parametrai pateikti 4 lentelėje.

Antrasis logistinės regresijos modelis sudarytas įtraukiant matematikos ir lietuvių kalbos PUPP reikšmes kaip kategorinį kintamąjį (kintamieji plačiau aprašomi 6.2 skyriuje ir jo poskyryje).

Modelis tiko duomenims, tikėtinumo santykio kriterijus $\chi^2 = 213379$, p-reikšmė < 0.001 . Kintamasis, kuris atspindi nekvalifikuotą darbą dirbančius tėčius (*Kvalif_tėtis_nekval*), modelyje yra nereikšmingas (Voldo kriterijaus p-reikšmė didesnė nei $\alpha = 0.05$). Jis iš modelio nebuvo šalinamas, tai viena iš kintamojo kategorijų. Determinacijos pseudokintamojo reikšmė yra didesnė nei pirmo modelio, $R^2 = 0.520$.

Modeliui sudaryta klasifikacinė lentelė (3 lentelė) parodo, kad teisingai klasifikuota 88.7 procentai stojusiųjų ir 85.5 procentai į AM nestojusiųjų studentų. Bendras teisingų

klasifikavimų procentas: 86.5 (4 lentelė).

7.2.3 3 modelis

Modelio parametrai pateikti 4 lentelėje.

Tračias daugialypės logistinės regresijos modelis sudarytas naudojant matematikos ir lietuvių kalbos PUPP NA reikšmes kiečiant 0 ir įtraukiant kintamuosius *Liet_PUPP_laikė* ir *Mat_PUPP_laikė* (kintamieji plačiau aprašomi 6.2 skyriuje ir jo poskyryje).

Modelis duomenims tiko, tikėtinumo santykio kriterijus $\chi^2 = 213063$, p-reikšmė < 0.001 . Kaip ir antrame modelyje, gavome kad kategorija, atspindinti nekvalifikuotą darbą dirbančius tėčius (*Kvalif_tėtis_nekval*) yra nereikšminga (Voldo kritrijaus p-reikšmė didesnė nei $\alpha = 0.05$). Determinacijos pseudokintamojo reikšmė yra tokia pati kaip ir modelio 2 - $R^2 = 0.520$.

Modelio tinkamumo vertinimui sudaryta klasifikacinė lentelė (3 lentelė) parodo, kad 82.4 procentai stojusiųjų ir 88.9 procentai nestojusiųjų į AM po bendrojo ugdymo mokyklos baigimo yra teisingai klasifikuoti. Bendras teisingų klasifikavimų procentas: 86.725 (4 lentelė).

7.2.4 1-3 modelių palyginimas

Modeliai lyginami naudojant modelio tinkamumą apibūdinančius rodiklius (4 lentelė). Toliau analizėje naudosime 2 modelį. 2 modelio bendras teisingai klasifikuotų reikšmių procentas nėra didžiausias, tačiau procentas labai ne daug skiriasi nuo kitų modelių. Determinacijos pseudokintamojo rodiklis yra toks pat kaip ir 3 modelio. AIC kriterijaus reikšmė 2 modelio mažiausia.

Modelio 2 išvados pateikiamos 8 skyriuje.

3 lentelė: 1-3 modelių klasifikacinės lentelės

	1 modelis		2 modelis		3 modelis	
Y	0	1	0	1	0	1
0	2365 (89.0%)	247 (18.4%)	2358 (85.5 %)	241 (11.3%)	2362 (88.9 %)	236 (17.6 %)
1	292 (11.0%)	1096 (81.6%)	399 (14.5%)	1102 (88.7%)	295 (11.1%)	1107 (82.4%)

4 lentelė: 1-3 modelių parametrai

	1 modelis			2 modelis			3 modelis		
	Koef.	SD	Galimybų santykis	Koef.	SD	Galimybų santykis	Koef.	SD	Galimybų santykis
Pajamos_Q3	-0.045***	(0.018)	0.956	-0.061***	(0.018)	0.941	-0.057**	(0.018)	0.865
Pajamos_Q2	-0.177*	(0.020)	0.838	-0.201***	(0.020)	0.818	-0.194***	(0.020)	0.720
Pajamos_Q1	-0.304***	(0.022)	0.738	-0.360***	(0.022)	0.698	-0.354***	(0.022)	0.593
Pajamos_Nedirbo samdomo darbo	-0.417***	(0.035)	0.659	-0.460***	(0.035)	0.631	-0.452***	(0.035)	0.529
Lytis_Mergaitė	0.266***	(0.012)	1.304	0.271***	(0.012)	1.311	0.261***	(0.012)	1.304
Gyvenvietė_Miestas	0.184***	(0.023)	1.202	0.166***	(0.023)	1.180	0.165***	(0.023)	1.170
Etninė_mažuma_taip	0.424***	(0.023)	1.529	0.412***	(0.023)	1.510	0.409***	(0.023)	1.510
Bent_vienas_aukštos_kval	0.304***	(0.037)	1.355	0.370***	(0.037)	1.448	0.380***	(0.037)	1.455
Kvalif_mama_nedirbo	0.103***	(0.030)	1.109	0.774***	(0.032)	2.168	0.765***	(0.032)	2.151
Kvalif_mama_nepriskirtos_kval	0.50***	(0.025)	1.663	0.396***	(0.025)	1.486	0.397***	(0.025)	1.400
Kvalif_mama_nekval	0.015	(0.038)	1.015	0.704***	(0.040)	2.021	0.697***	(0.040)	1.951
Kvalif_mama_žemos_kval	0.104**	(0.032)	1.109	0.808***	(0.035)	2.244	0.807***	(0.035)	2.160
Kvalif_mama_aukšt_kval	-0.046	(0.042)	1.047	0.724***	(0.044)	2.063	0.713***	(0.044)	1.906
Kvalif_tėtis_nedirbo	0.116***	(0.026)	1.123	0.316***	(0.026)	1.372	0.319***	(0.026)	1.377
Kvalif_tėtis_nepriskirtos_kval	0.342***	(0.021)	1.407	0.253***	(0.021)	1.288	0.255***	(0.022)	1.203
Kvalif_tėtis_nekval	-0.106***	(0.040)	0.900	0.070	(0.040)	1.073	0.071	(0.040)	1.041
Kvalif_tėtis_žemos_kval	0.165***	(0.027)	1.088	0.354***	(0.027)	1.424	0.361***	(0.027)	1.357
Kvalif_tėtis_aukštos_kval	0.007	(0.033)	1.007	0.193***	(0.033)	1.213	0.191***	(0.033)	1.117
Liet_PUPP_vid	-0.000	(0.006)	1.000						
Mat_PUPP_vid	-0.015***	(0.004)	0.985						
Liet_PUPP_kat_nepatenkinamas				0.789***	(0.091)	2.200			
Liet_PUPP_kat_patenkinamas				2.019***	(0.064)	7.5311			
Liet_PUPP_kat_pagrindinis				2.085***	(0.063)	8.048			
Liet_PUPP_kat_aukštesnysis				1.975***	(0.067)	7.210			
Mat_PUPP_kat_nepatenkinamas				-3.155***	(0.066)	0.043			
Mat_PUPP_kat_patenkinamas				-3.186***	(0.064)	0.041			
Mat_PUPP_kat_pagrindinis				-3.216***	(0.065)	0.038			
Mat_PUPP_kat_aukštesnysis				-3.406***	(0.068)	0.033			
Mat_PUPP_0							-0.041***	(0.004)	0.959
Liet_PUPP_0							0.023***	(0.006)	1.024
Mat_PUPP_laikė							-2.717***	(0.056)	0.067
Liet_PUPP_laikė							1.601***	(0.060)	4.987
Egz_suma	1.173***	(0.005)	3.232	1.122***	(0.005)	3.071	1.131***	(0.005)	3.087
Konstanta	-3.771***	(0.052)	0.023	-3.241***	(0.043)	0.035	-3.367***	(0.043)	0.042
Tikėtinumo santykio kriterijus		207714***			213379***			213063***	
Determinacijos pseudokoefficientas (Makfaden)		0.507			0.520			0.520	
Akaike informacinis kriterijus (AIC)		202302			196648			196957	
Teisingai klasifikuotos reikmės		86.525%			86.5%			86.725%	

Pastaba:

*p<0.1; **p<0.05; ***p<0.01; ****p<0.001

7.3 Daugialypė atitikimo analizė

Į 7.2 skyriuje apibrėžtus modelius įtraukta nuo 21 iki 27 regresorių, dalis jų - tarpusavyje stipriai susijusių požymių, pavyzdžiui, SES apibūdinantys kintamieji: *pajamos*, *kvalif_mama*, *kvalif_tėtis* ar studijų pajėgumą atspindintys kintamieji: *mat_PUPP_kat*, *liet_PUP_kat*. Taikant statistinės analizės metodą - DAA, iš tarpusavyje susijusių kintamųjų sukurtos naujos komponentės paaiškinančios duomenų variaciją. Kadangi tik dalis komponentių įtraukta į naujus modelius, taip sumažintas nepriklausomų kintamųjų daugialypėje logistinėje regresijoje skaičius.

Toliau esančiuose poskyriuose, naudojant DAA metodą, sukurtos komponentės, kurios atspindi SES bei studijų pajėgumą.

7.3.1 SES kintamasis

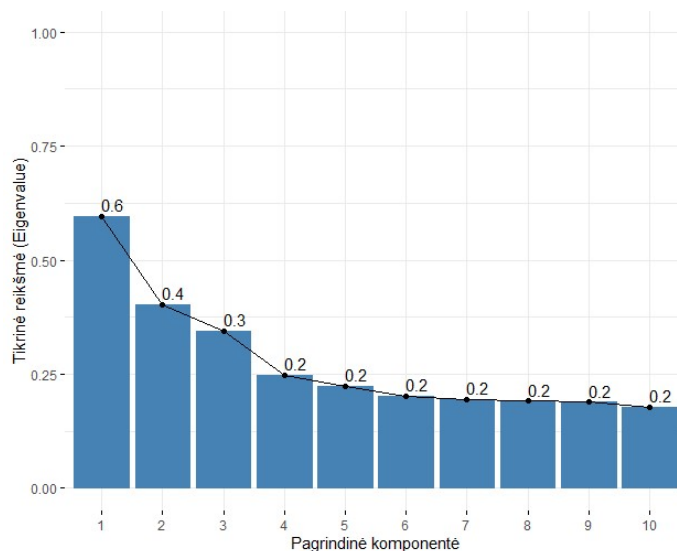
Kuriant SES atspindintį kintamąjį, į DAA įtraukta:

- *Pajamos*
- *Kvalif_mama*
- *Kvalif_tėtis*
- *Bent_vienas_aukšt_kval*
- *Gyvenvietė*

Daugialypės atitikimo analizės sukurtų komponentių statistikos pateikiamos 5 lentelėje. Tikrinės reikšmės nusako ašių, einančių per duomenų taškus ilgį, o dispersija - kiek procentų bendros informacijos išlaiko kiekviena komponentė [18].

Įvertinti kaip staigiai keičiasi tikrinų reikšmių skirtumai sprendžiant iš 6 pav. - sunku. Skirtumai sumažėja ties 4 komponente. Komponentų, pradedant nuo 4, tikrinės reikšmės vienodos - 0.2 (6 pav.). Remiantis tikrinių reikšmių skirtumų sumažėjimu į 7.3.3, 7.3.4 skyriuose aprašomus logistinės regresijos modelius įtrauksime tik pirmas 3 komponentes, kurios bendrai paaiškina 41.986% variacijos (5 lentelė). Paaiškinamumo procentas palyginti nedidelis, siekiant jį padidinti į 7.3.3 skyriuje aprašomą modelį buvo įtraukta ketvirtoji DAA metodu sukurto SES kintamojo komponentė (bendras paaiškinamumas padidėja iki 49.647% (5 lentelė)). Ji buvo nereikšmingas kintamasis (Voldo kriterijaus p-reikšmė didesnė už $\alpha = 0.05$), todėl iš galutinio modelio pašalinta.

6 pav.: Tikrinių reikšmių grafikas (angl. Scree Plot)



7.3.2 Studijų pajėgumo kintamasis

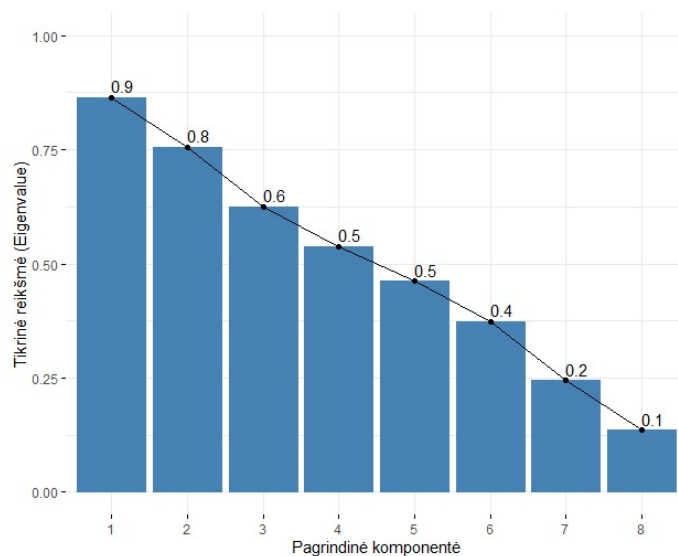
Kuriant studijų pajėgumą atspindintį kintamąjį, į DAA įtraukta:

- *Mat_PUPP_kat*
- *Liet_PUP_kat*

5 lentelėje pateikta daugialypės atitikimo analizės sukurtų komponentių statistikos. Tikrinių reikšmių kitimo greitis yra tolydus, nėra staigaus reikšmių sumažėjimo (7 pav.). Į penktą logistinės regresijos modelį įtraukta tik vieną principinę komponentę, kuri paaiškina 21.617% dispersijos (5 lentelė). Dispersijos paaiškinimo procentas nedidelis. Buvo pabandyta įtraukti antrąją principinę komponentę - dispersijos paaiškinamumas padidėja iki 40.517% (5 lentelė). Tačiau modelio teisingai klasifikuotų reikšmių procentas padidėja tik per 0.05%.

DAA analizė buvo taikoma siekiant sumažinti regresorių, įtraukiamų į modelius, skaičių. Studijų pajėgumą atspindinčio kintamojo antroji komponentė pagerina paaiškinamą dispersijos procentą, tačiau teisingai klasifikuotų reikšmių procentas keičiasi labai mažai, todėl ji, į galutinį 7.3.4 skyriuje aprašomą modelį nėra įtraukiama.

7 pav.: Tikrinių reikšmių grafikas (angl. Scree Plot)



5 lentelė: DAA rezultatai

Komponentės	Tikrinė reikšmė (angl. - eigenvalue)	Dispersija procentais	Dispersijų suma procentais
Komp.1_SES	0.605	18.894	18.894
Komp.2_SES	0.408	12.759	31.653
Komp.3_SES	0.331	10.333	41.986
Komp.4_SES	0.245	7.661	49.647
Komp.5_SES	0.224	6.986	56.634
Komp.6_SES	0.201	6.296	62.929
Komp.7_SES	0.194	6.051	68.980
Komp.8_SES	0.192	5.986	74.966
Komp.9_SES	0.189	5.903	80.869
Komp.10_SES	0.178	5.569	86.438
Komp.11_SES	0.143	4.468	90.906
Komp.12_SES	0.132	4.136	95.042
Komp.13_SES	0.098	3.076	98.118
Komp.14_SES	0.027	0.852	98.970
Komp.15_SES	0.020	0.616	99.586
Komp.16_SES	0.013	0.414	100
Komp.1_PUPP	0.865	21.617	21.617
Komp.2_PUPP	0.756	18.900	40.517
Komp.3_PUPP	0.626	15.645	56.161
Komp.4_PUPP	0.539	13.463	69.624
Komp.5_PUPP	0.461	11.537	81.161
Komp.6_PUPP	0.374	9.355	90.517
Komp.7_PUPP	0.244	6.100	96.617
Komp.8_PUPP	0.135	3.383	100

7.3.3 4 modelis

Modelio parametrai pateikti 6 lentelėje.

Ketvirtas daugialypės logistinės regresijos modelis sudarytas su sukurtu SES kintamuoju (plačiau: 7.3.1 skyriuje), matematikos ir lietuvių kalbos PUPP reikšmės įtraukti kaip faktoriai, remiantis 7.2.4 skyrelio išvadomis. Visi modelio nepriklausomi kintamieji statistiškai reikšmingi. Modelis tiko duomenims, tikėtinumo santykio kriterijus $\chi^2 = 212824$, p-reikšmė < 0.001 . Determinacijos pseudokintamojo reikšmė didelė - $R^2 = 0.520$.

Modelio tinkamumo vertinimo klasifikacinė lentelė (7 lentelė) parodo, kad 82.4 procentai stojusiųjų ir 88.7 procentai nestojusiųjų į aukštąsias mokyklas Lietuvoje po BU mokyklos baigimo yra teisingai klasifikuoti. Bendras teisingų klasifikavimų procentas: 86.575 (6 lentelė).

7.3.4 5 modelis

Modelio parametrai pateikti 6 lentelėje.

Paskutinis modelis sukurtas naudojant kintamuosius, kurie sukurti taikant DAA (plačiau 7.3.1, 7.3.2 poskyriuose). Logistinėje regresijoje liko tik 7 nepriklausomi kintamieji, kai tuo tarpu 2 modelyje jų buvo 27 (su 2 modeliu lyginame remiantis 7.2.4 skyriumi). Regresorių skaičius ženkliai sumažėjo, tačiau prarasta tik dalis duomenyse esamos informacijos. Visi logistinės regresijos nepriklausomi kintamieji reikšmingi. Modelis išliko tinkamas duomenims, tikėtinumo santykio kriterijus $\chi^2 = 209572$, p-reikšmė < 0.001 . Determinacijos pseudokintamojo reikšmė mažesnė nei 4 modelio - $R^2 = 0.511$.

Remiantis 7 lentele, teisingai klasifikuota 89.0 procentai nestojusiųjų ir 81.8 procentai stojusiųjų į AM abiturientų. Bendras teisingų klasifikavimų procentas: 86.6 (6 lentelė).

7.3.5 4-5 modelių palyginimas

Visi modelių parametrai pateikti 6 lentelėje.

Abiejuose logistinės regresijos modeliuose esantys nepriklausomi kintamieji yra statistiškai reikšmingi (Voldo kriterijaus p-reikšmės mažesnės už reikšmingumo lygmens $\alpha = 0.05$ reikšmę). Abu modeliai tinka duomenims, tikėtinumo santykio kriterijaus p-reikšmės mažesnės už $\alpha = 0.05$. Determinacijos pseudokoefficientas didesnis ir AIC kriterijaus mažesnis 4 modelio. Tačiau remiantis teisingai klasifikuotų reikšmių procentais, 5 modelis geriau tei-

singai klasifikuoja reikšmes. 5 modelyje ženkliai sumažėjo regresorių skaičius, o teisingai klasifikuotų reikšmių procentas išlieka aukštas. Toliau analizėje naudosime 5 modelį.

7.4 2 modelio ir 5 modelio palyginimas

Remiantis 7.2.4 poskyrio išvadomis, analizėje naudojamas 2 modelis. Remiantis 7.3.5 poskyriu, pritaikius DAA, analizėje naudojamas 5 modelis. Pagal 4 ir 6 lentelėse pateiktus logistinio modelio patikimumo parametrus, 5 modelis laikomas geresniu už 2 modelį. 5 modelio teisingai klasifikuotų reikšmių skaičius didesnis, nors ir AIC bei determinacijos pseudokoeficiento reikšmės mažesnės.

5 modelio išvados pateikiamos 8 skyriuje.

6 lentelė: 4-5 modelių parametrai

	4 modelis			5 modelis		
	Koef.	S.d	Galimybių santykis	Koef.	S.d	Galimybių santykis
Komp.1_SES	0.710***	(0.012)	2.034	0.620***	(0.011)	1.858
Komp.2_SES	0.262***	(0.011)	1.299	0.308***	(0.011)	1.360
Komp.3_SES	0.182***	(0.013)	1.306	0.139***	(0.013)	1.199
Komp.1_PUPP				-0.508***	(0.011)	0.602
lytis_Mergaitė	0.265***	(0.012)	1.303	0.264***	(0.011)	1.302
etninė_mažuma_taip	0.402***	(0.023)	1.494	0.437***	(0.022)	1.548
Liet_PUPP_kat_nepatenkinamas	0.815***	(0.091)	2.259			
Liet_PUPP_kat_patenkinamas	2.030***	(0.065)	7.616			
Liet_PUPP_kat_pagrindinis	2.096***	(0.064)	8.136			
Liet_PUPP_kat_aukštesnysis	1.987***	(0.068)	7.291			
Mat_PUPP_kat_nepatenkinamas	-3.122***	(0.066)	0.044			
Mat_PUPP_kat_patenkinamas	-3.152***	(0.065)	0.043			
Mat_PUPP_kat_pagrindinis	-3.228***	(0.065)	0.040			
Mat_PUPP_kat_aukštesnysis	-3.374***	(0.068)	0.034			
Egz_suma	1.130***	(0.005)	3.096	1.217***	(0.004)	3.376
Kontstanta	-2.748***	(0.018)	0.064	-3.650***	(0.014)	0.026
Tikėtinumo santykio kriterijus		212824***			209572***	
Determinacijos pseudokoeficientas (Makfaden)		0.520			0.511	
Akaike informacinis kriterijus (AIC)		197178			200415	
Teisingai klasifikuotos reikšmės		86.575%			86.6%	

Pataba:

p<0.1, *p<0.05; **p<0.01; ***p<0.001

7 lentelė: 4-5 modelių klasifikacinės lentelės

Y	4 modelis		5 modelis	
	0	1	0	1
0	2357 (88.7%)	237(17.6%)	2366 (89.0 %)	245 (18.2%)
1	300 (11.3%)	1106 (82.4%)	291 (11.0%)	1098 (81.8%)

8 Išvados ir rekomendacijos

Analizėje, naudojant 2013-2014 - 2019-2020 mokslo metų duomenis sudaryti penki daugialypės logistinės regresijos modeliai. Trys iš jų sudaryti naudojant minimaliai modifikuotus duomenis. Likę du sudaryti su duomenimis, kuriems buvo pritaikyta daugialypės atitikimo analizė, taip siekiant sumažinti regresorių skaičių modeliuose. Į analizę yra įtraukti tik bendrojo ugdymo mokyklų moksleiviai. Stojančiais į AM laikomi tie, kurie tęsia studijas Lietuvos aukštojoje mokykloje.

Iš 2 modelio gautų rezultatų (4 lentelė) negalime patvirtinti, kad žemesnio SES moksleiviai linkę rečiau tęsti studijas aukštojoje mokykloje, kandangi visi koeficientai prie *pajamų* kintamojo įgauna neigiamas reikšmes. Jei abu vaiko tėvai nedirbo, tai tikimybė, kad vaikas tęs studijas aukštojoje mokykloje mažėja $\frac{1}{0.631} = 1.585$ karto (lyginant su $\alpha = 0.05$ santykis statistiškai reikšmingai skiriasi nuo 1) (4 lentelė), o jei vaiko tėvai gauna pajamas, kurios patenka į trečią pajamų kvartilį, tai tikimybė, kad vaikas tęs studijas AM mažėja $\frac{1}{0.941} = 1.063$ karto (lyginant su $\alpha = 0.05$ santykis statistiškai reikšmingai skiriasi nuo 1) (4 lentelė). Be to, jei bent vienas iš tėvų dirbą aukštos kvalifikacijos darbą, tikimybė, kad vaikas studijuos aukštojoje mokykloje didėja 1.355 karto (lyginant su $\alpha = 0.05$ santykis statistiškai reikšmingai skiriasi nuo 1) (4 lentelė).

Atlikus daugialypę atitikimo analizę ir sumažinus į 5 modelį įtraukiamų regresorių skaičių naudojant gautas komponentes daromos išvados:

1. Mergaičių šansai stoti/įstoti į AM yra 1.302 karto didesni nei berniukų (6 lentelė).
2. Etninei mažumai priklausančių moksleivių šansai stoti/įstoti į AM 1.548 karto didesni, nei etninei mažumai nepriklausančių jaunuolių (6 lentelė).
3. Papildomas laikytas valstybinis brandos egzaminas padidina galimybę, kad vaikas studijuos aukštojoje mokykloje 3.376 karto (6 lentelė). Be to, tai daugiausia įtakos stojimams į aukštąsias mokyklas turintis kintamasis.
4. Svarbiausia, kad daugialypės atitikimo analizės pagalba sukurtų komponentių įtraukimas į modelį padėjo pagerinti bendrą teisingai klasifikuojamų reikšmių procentą. Lyginant 2 (4 lentelė) ir 5 (6 lentelė) modelių teisingai klasifikuotų reikšmių procentą, penkto modelio bendras teisingai klasifikuotų reikšmių procentas yra didesnis puse

procento. Darome išvadas, kad pritaikius daugialypę atitikimo analizę ir į modelį įtraukus žymiai mažiau priklausomų kintamųjų tik dalis informacijos buvo prarasta, tačiau modelio patikimumas padidėjo.

DAA metodo taikymas gali būti naudingas ateityje atliekant panašų tyrimą, kuriame turima mažiau stebinių. Metodo pagalba yra išvengiama problema, susijusi su mažu stebinių kiekiu, bet dideliu skirtingų požymių skaičiumi duomenyse. Be to, metodo pagalba sukurtos komponentės leistų stebėti moksleivius, apimant visus, juos apibūdinančius kintamuosius.

Tolimesni šio darbo analizės aspektai galėtų būti:

1. Darbe buvo žiūrima tik į 10-os klasės PUPP rezultatus, norint išsiaiškinti bendrojo ugdymo mokyklose gaunamų rezultatų įtaką stojimams, naudinga būtų pažiūrėti ir į valstybinių brandos egzaminų ar į skirtingų pamokų gaunamų pažymių vidurkių rezultatus.
2. Analizė būtų išplėsta, jei žiūrėtume ne tik į bendras stojimo į AM tendencijas. Tikslinant gautus rezultatus, būtų galima sudarinėti modelius, kurie leistų įvertinti stojimus į AM pagal studijų programas, aukštosios mokyklos tipą.
3. Į analizę buvo įtraukti tik BU mokyklų 10-os klasės moksleiviai. Norint ateityje išplėsti tyrimo apimtį, į tiriamųjų imtį būtų galima įtraukti ir PU mokyklų studentus.
4. Duomenys atspindėjo stojimus tik į Lietuvoje esančias aukštąsias mokyklas. Tam, kad tyrimas būtų dar išsamesnis, būtų galima pažiūrėti ir stojimus į užsienyje esančias mokyklas.

Literatūra

- [1] https://lt.wikipedia.org/wiki/Auk%C5%A1tasis_i%C5%A1silavinimas
- [2] https://ec.europa.eu/education/policies/higher-education/about-higher-education-policy_lt
- [3] Vyriausybės strateginės analizės centras (Strata) 2018 m. *Aukštojo mokslo prieinamumas: ar skirtigo SES vaikams sudarome lygias galimybes įgyti išsilavinimą*
<https://strata.gov.lt/images/tyrimai/am-prieinamumas-2018.pdf>
- [4] Mokslo ir studijų stebėsenos ir analizės centras (Mosta) 2014 m, *Studijų pasirinkimo motyvai ir pasirengimas studijoms*
https://strata.gov.lt/images/leidiniai/Studiju_pasirinkimo_motyvai_TRUMPRASTIS.pdf
- [5] G.James, D.Witten, T.Hastie, R.Tibshirani *"An Introduction to Statistical Learning"*. Springer, 2013 (Corrected at 8th printing 2017).
- [6] V.Čekanavičius, G.Murauskas *Logistinė regresija* <http://stat.vadoveliai.lt/files/LogRegSAS.pdf>
- [7] V. Kazakevičius *Kokybinių duomenų analizė* <https://klevas.mif.vu.lt/vytas/stat/KDA-suR.pdf>
- [8] G.Dzemyda, O.Kurasova, J.Žilinskas *Daugiamatčių duomenų vizualizavimo metodai*. Vilnius, 2008 <http://web.vu.lt/mii/j.zilinskas/DzemydaKurasovaZilinskasDDVM.pdf>
- [9] https://lt.wikipedia.org/wiki/Socioekonomin%C4%97_pad%C4%97tis
- [10] https://lt.wikipedia.org/wiki/Auk%C5%A1toji_mokykla
- [11] <http://www.profesijuklasifikatorius.lt/?q=lt/node/13440>
- [12] V.Čekanavičius, G.Murauskas *Taikomoji regresinė analizė socialiniuose tyrimuose*
<http://www.statistika.mif.vu.lt/wp-content/uploads/2014/04/regresine-analize.pdf>
- [13] <https://bakalauras.lamabpo.lt/stojanciuju-skaicius-ir-ju-pasirengimas/>
- [14] <https://www.kurstoti.lt/s/1394/is-mokyklos-suolo-i-bemoksliu-ir-bedarbiu-gretas-kodel>

- [15] J. Ma, M. Pender, M. Welch *Education Pays 2019 THE BENEFITS OF HIGHER EDUCATION FOR INDIVIDUALS AND SOCIETY*, 2020
- [16] M.Kuhn, K. Johnson *Applied predictive modeling*. Springer, 2013
- [17] [https://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))
- [18] V.Čekanavičius, G.Murauskas *Statistika ir jos taikymas II*. Vilniu, 2002
- [19] M.Radavičius *Regresiniai Modeliai ir Jų Tyrimas*. Vilnius, 2009
https://www.academia.edu/9626159/Regresiniai_Modeliai_ir_J%C5%B3_Tyrimas_4_Did%C5%BEiausio_tik%C4%97tinumo_teorijos_pagrindai_26_4.1_Tik%C4%97tinumo_funkcija_ir_Fi%C5%A1erio_informacija
- [20] J. Pagès *Multiple Factor Analysis by Example Using R*. Chapman Hall/CRC The R Series, London, 2015

9 Priedai

9.1 Dešimtbalė vertinimo skalė pagrindiniame ir viduriniame ugdyme

8 lentelė: Dešimtbalė vertinimo skalė pagrindiniame ir viduriniame ugdyme

Pasiekimų lygis	Balai
Aukštesnysis	10 (dešimt)
	9(devyni)
Pagrindinis	8 (aštuoni)
	7 (septyni)
	6 (šeši)
Patenkinamas	5 (penki)
	4 (keturi)
Nepatenkinamas	3 (trys)
	2 (du)
	1(vienas)

9.2 R programos kodas

```
#Nustatome lietuviškoms raidėms pritaikytas vietas (angl.- locale)
```

```
Sys.getlocale()
```

```
Sys.getlocale("LC_CTYPE")
```

```
Sys.setlocale("LC_CTYPE", "Lithuanian_Lithuania.1257")
```

```
#Darbe naudojamos bibliotekos:
```

```
library(dplyr)
```

```
library(reshape2)
```

```
library(ggplot2)
```

```
library(scales)
```

```
library(data.table)
```

```

library(DescTools)
library(mice)
library(VIM)
library(ggpubr)
library(gplots)
library(FactoMineR)
library(factoextra)
library(bios2mds)
library(psych)
library(GPArotation)
library(REdaS)

# paletės
pagr_palette=c("blue", "grey", "gold",
               "black","red")

#Duomenų nuskaitymas:
data=read.csv("Duomenys.csv", header=TRUE, sep=",",
              row.names=NULL, na.strings=c("", "NA"), stringsAsFactors=FALSE)

#Duomenų transformacijos:
data=data %>% mutate(etnine_mazuma=ifelse(is.na(kalba),NA,ifelse(kalba=="lietuvių","ne","taip")))
summary(as.factor(data$etnine_mazuma))
(370/(length(data$etnine_mazuma))*100
data=data %>% mutate(gyvenvietė=ifelse(vietovės_tipas=="Kaimas","Kaimas","Miestas"))

data=data %>%
  mutate(bu_baig_metai=ifelse(is.na(metai),NA,as.numeric(substr(metai, 6,9))),
         metai_istojimo2=ifelse(is.na(metai_istojimo),NA,
                                as.numeric(substr(metai_istojimo,1,4))),
         fin_tipas_bi=ifelse(is.na(fin_tipas),"Neįstojo",
                              ifelse(fin_tipas=="Mokantys uz mokslą","VNF","VF"))),

```

```

istojo2=ifelse(is.na(metai_istojimo),0,1),
istojo_tri=ifelse(istojo2==0, "neistojo",
                  ifelse(istojo2==1&is.na(bu_baig_metai),"istojo,
                        nežinoma kada",
                        ifelse(metai_istojimo2==bu_baig_metai,
                              "istojo be pertraukos","istojo po pertraukos"))),
istojo=ifelse(istojo_tri=="istojo be pertraukos",1,0),
egz_suma=ifelse(is.na(egz_sum_VirM),0,egz_sum_VirM))

```

```

data=data %>%

```

```

mutate(kvalif_mama=ifelse(is.na(mama)&is.na(paj_m), "neturi",
                        ifelse(is.na(lpk_m)&is.na(paj_m)&!is.na(mama),"nedirbo",
                        ifelse(is.na(lpk_m)&!is.na(paj_m),"nepriskirtos_kval",
                        ifelse(substr(lpk_m,1,1)<4,"aukstos_kval",
                        ifelse(substr(lpk_m,1,1)>=4&substr(lpk_m,1,1)<9, "zemos_kval",
                        ifelse(substr(lpk_m,1,1)==9,"nekval","ERROR"))))))),
kvalif_tetis=ifelse(is.na(tetis)&is.na(paj_t), "neturi",
                    ifelse(is.na(lpk_t)&is.na(paj_t)&!is.na(tetis),"nedirbo",
                    ifelse(is.na(lpk_t)&!is.na(paj_t),"nepriskirtos_kval",
                    ifelse(substr(lpk_t,1,1)<4,"aukstos_kval",
                    ifelse(substr(lpk_t,1,1)>=4&substr(lpk_t,1,1)<9, "zemos_kval",
                    ifelse(substr(lpk_t,1,1)==9,"nekval","ERROR"))))))),
kvalif_tetis=factor(kvalif_tetis,
levels=c("neturi","nedirbo","nepriskirtos_kval", "nekval","zemos_kval",
"aukstos_kval")),
kvalif_mama=factor(kvalif_mama,
levels=c("neturi","nedirbo","nepriskirtos_kval", "nekval",
"zemos_kval","aukstos_kval")),
bent_vienas_aukstos_kval=
ifelse(kvalif_tetis=="aukstos_kval"|kvalif_mama=="aukstos_kval",1,0))

```

```
#Panaikiname Bendrojo ugdymo mokyklose besimokančius mokinius bei etninės
mažumos NA reikšmes:
```

```
prop.table(table(data$pm_bu))
data2<-data %>% filter( pm_bu == "BU", is.na(etnine_mazuma)== FALSE)
```

```
#Sukuriami pajamų kintamieji: bendros tėvų pajamos ir jų kvantiliai:
```

```
data2=data2 %>%
  mutate(paj_t0=ifelse(is.na(paj_t),0, paj_t),
         paj_m0=ifelse(is.na(paj_m),0,paj_m),
         paj_tevu_0=paj_t0+paj_m0,
         paj_tevu=ifelse(paj_tevu_0==0,NA,paj_tevu_0),
         pajamos=ifelse(is.na(paj_tevu),
                        "Nedirbo samdomo darbo", ntile(paj_tevu,4)),
         pajamos=factor(pajamos,
                        levels=c('4','3','2','1',"Nedirbo samdomo darbo")))%>%
  ungroup()
summary(data2$paj_tevu)
```

```
#PM ir BU įstojimai po mokyklos baigimo i AM
```

```
tab0<-data%>%
  group_by(istojo, pm_bu)%>%
  dplyr::summarise(sk=n())%>%
  group_by(pm_bu)%>%
  mutate(dalis=round(sk/sum(sk),2))%>%
  print()
```

```
#Lietuviu ir matematikos PUPP NA procentais
```

```
summary(data2$matematika_PUPP)
118741/length(data2$matematika_PUPP)
summary(data2$lietuviu_PUPP)
```

```
168496 /length(data2$lietuviu_PUPP)
```

```
#Vizualinės analizės grafikai ir skaičiavimai:
```

```
#Mokinių stojimas į AM:
```

```
tab=data2 %>%
```

```
  group_by(metai_10oj_numb,pajamos,istojo) %>%
```

```
  dplyr::summarise(sk=n()) %>%
```

```
  group_by(metai_10oj_numb, pajamos) %>%
```

```
  mutate(dalis=round(sk/sum(sk),2)) %>%
```

```
  filter(istojo==1& metai_10oj_numb<2018)%>%
```

```
  print()
```

```
tab2= data2 %>% group_by( metai_10oj_numb, istojo ) %>%
```

```
  dplyr:: summarise(sk=n()) %>%
```

```
  group_by(metai_10oj_numb) %>%
```

```
  mutate(dalis=round(sk/sum(sk),2)) %>%
```

```
  filter(istojo ==1 & metai_10oj_numb <2018)%>%
```

```
  print()
```

```
A<-ggplot(tab,aes(x = metai_10oj_numb, y = dalis,color=as.factor(pajamos)))+  
geom_line()+ scale_x_continuous(name="Metai, kai mokinys buvo 10-oje klasėje")+  
  scale_y_continuous(name="Įstojusiųjų į AM dalis", limits = c(0, 0.8))+  
  scale_color_manual( values = pagr_palette, name="Tėvų pajamų kvartilis:",  
                      labels=c(" Q4", " Q3"," Q2"," Q1","Nedirbo"))+  
  theme(axis.text=element_text(size=9),  
        axis.title=element_text(size=9))
```

```
B<-ggplot(data=tab2, aes(x=metai_10oj_numb, y=dalis, group=1)) +  
  geom_line()+geom_point()+scale_x_continuous(name="Metai,  
  kai mokinys buvo 10-oje klasėje")+  
  scale_y_continuous(name="Įstojusiųjų į AM dalis", limits = c(0.4, 0.5))  
+theme(axis.text=element_text(size=9),axis.title=element_text(size=9))
```

```

gridExtra::grid.arrange(B, A, nrow=1)

vidutinis_skirtumas<-mean(c(0.75-0.25, 0.75-0.27,
0.73-0.26, 0.70-0.25, 0.66 - 0.24, 0.62-0.22, 0.62 - 0.23))
vidutinis_skirtumas

#Mokinių stojimas į AM pagal finansavimo pobūdį
tab3=data2 %>%
  group_by(pajamos, fin.tipas_bi) %>%
  dplyr::summarise(sk=n()) %>%
  group_by(pajamos) %>%
  mutate(dalis=round(sk/sum(sk),3)) %>%
  print()

ggplot(tab3, aes(pajamos, dalis, fill=fin.tipas_bi))+
  geom_bar(stat="identity",position="stack")+scale_fill_manual("Fin. tipas",
  values = c("Neįstojo" = "gold", "VF" = "darkorange", "VNF" = "brown"))+
  scale_x_discrete(name="Tėvų pajamų kartilis",
  labels=c(" Q4", " Q3"," Q2"," Q1","Nedirbo"))+
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12))+
  scale_y_continuous(name="Pasiskirstymo dalis")

#Mokinių stojimas į AM pagal PUPP pasiekimus
tab4=data2 %>%
  group_by(matematika_PUPP,lytis, fin.tipas_bi) %>%
  dplyr::summarise(sk=n()) %>%
  group_by(matematika_PUPP,lytis) %>%
  mutate(dalis=round(sk/sum(sk),3)) %>%
  print()
tab4$dalis

```

```

C<-ggplot(tab4, aes(matematika_PUPP, dalis, fill=fin.tipas_bi))+
  geom_bar(stat="identity",position="stack")+ facet_wrap(~lytis)+
  scale_fill_manual("Fin. tipas",
    values = c("Neįstojo" = "gold", "VF" = "darkorange", "VNF" = "brown"))+
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12))+
  scale_y_continuous(name="Pasiskirstymo dalis")

tab5=data2%>%
  group_by(lietuviu_PUPP,lytis, fin.tipas_bi) %>%
  dplyr::summarise(sk=n()) %>%
  group_by(lietuviu_PUPP,lytis) %>%
  mutate(dalis=round(sk/sum(sk),3)) %>% filter(fin.tipas_bi=="Neįstojo") %>%
  print()
tab5$dalis

D<-ggplot(tab5, aes(lietuviu_PUPP, dalis, fill=fin.tipas_bi))+
  geom_bar(stat="identity",position="stack")+ facet_wrap(~lytis)+
  scale_fill_manual("Fin. tipas",
    values = c("Neįstojo" = "gold", "VF" = "darkorange", "VNF" = "brown"))+
  theme(axis.text=element_text(size=12),
    axis.title=element_text(size=12))+scale_y_continuous(name="Pasiskirstymo dalis")

gridExtra::grid.arrange(D, C, nrow=1)

#berniukai-mergaitės
Neistojo.mat<-mean(c(0.969-0.927,0.934-0.857,0.858- 0.745, 0.832-0.715, 0.694-
0.576, 0.6- 0.511,0.474 - 0.431, 0.391- 0.380,0.329-0.316, 0.303-0.278 ))
VF.mat<- mean(c(0-0.004, 0.009-0.017, 0.031-0.046, 0.045- 0.066, 0.125-0.159,
0.217 - 0.237, 0.352 - 0.345, 0.466-0.432, 0.566- 0.538,0.640-0.625 ))
VNF.mat<-mean(c(0.031-0.069, 0.057-0.126,0.111-0.209, 0.123-0.218,
0.177-0.269, 0.183-0.252, 0.174-0.225, 0.143-0.188, 0.105-0.147,0.057- 0.097))
VNF.Lt<- mean(c(0.003-0,0.013-0.014, 0.050-0.046, 0.096-0.102, 0.182-0.211,

```

```

0.225-0.317, 0.218-0.346, 0.177-0.296 ,0.126-0.216, 0.085-0.127, 0.101-0.140))
VF.Lt<- mean(c(0-0, 0.001-0,0.014-0.001, 0.044-0.011, 0.122-0.044, 0.273-0.133,
0.438-0.289, 0.569-0.453, 0.670-0.588 ,0.689-0.690, 0.127-0.159 ))
Neistojo.LT<-mean(c( 0.997-1.000, 0.985-0.986, 0.935-0.953, 0.861-0.887,
0.696-0.745 ,0.502-0.550, 0.344-0.365, 0.253-0.251, 0.203-0.196,
0.226-0.183, 0.771-0.701))

```

```

((VNF.mat+VNF.Lt)/2)*100

```

```

((VF.mat+VF.Lt)/2)*100

```

```

summary(as.factor(data2$metai_istojimo))

```

```

#Del duomenų netikslumo, panaikinamos 6 reiksmės 2012-2013 mokslo
metais stojančių moksleivių

```

```

data2<-data2 %>% mutate(istojo_kada_naikinimui= ifelse(is.na(metai_istojimo),
0, metai_istojimo)) #NA priskiriama 0, nežinome kada mokinsys stojo į AM,
jei iš vis stojo.

```

```

prop.table(table(data2$istojo))

```

```

table(data2$istojo_kada_naikinimui)

```

```

data2<-data2%>% filter(istojo_kada_naikinimui != "2012-2013")

```

```

data2$metai_istojimo<-data2$istojo_kada_naikinimui

```

```

#Matematikos ir lietuvių kalbos NA reiksmės keičiamos 0:

```

```

data2<-data2 %>% mutate(mat_PUPP_0 = ifelse(is.na(data2$matematika_PUPP),0,
data2$matematika_PUPP),

```

```

      liet_PUPP_0 = ifelse(is.na(data2$lietuviu_PUPP),0,
data2$lietuviu_PUPP))

```

```

#sukuriamas kintamasis, ar mokinsys laikė ar ne PUPP egzaminą:

```

```

data2<-data2 %>% mutate(mat_PUPP_laike = ifelse(mat_PUPP_0==0,0, 1),

```

```

      liet_PUPP_laike = ifelse(liet_PUPP_0 ==0,0, 1))

```

```

#Matematikos ir lietuvių kalbos NA reiksmės keičiamos į vidurkius:

```

```

data2<-data2 %>% mutate(mat_PUPP_vid = ifelse(is.na(data2$matematika_PUPP),
mean(data2$matematika_PUPP, na.rm=TRUE), data2$matematika_PUPP),
      liet_PUPP_vid = ifelse(is.na(data2$lietuviu_PUPP),
      mean(data2$lietuviu_PUPP, na.rm=TRUE), data2$lietuviu_PUPP))

#Matematikos ir lietuvių kalbos reikšmes keičiamos į kategorijas:
data2<-data2 %>% mutate(mat_PUPP_kat = ifelse(is.na(data2$matematika_PUPP),
"nelaike", ifelse(data2$matematika_PUPP ==1 | data2$matematika_PUPP ==2|
data2$matematika_PUPP ==3, "nepatenkinamas", ifelse(data2$matematika_PUPP ==4 |
data2$matematika_PUPP ==5, "patenkinamas", ifelse(data2$matematika_PUPP ==6 |
data2$matematika_PUPP ==7 | data2$matematika_PUPP == 8, "pagrindinis",
ifelse(data2$matematika_PUPP==9 | data2$matematika_PUPP==10, "aukstesnysis",
"x"))))))))
data2$mat_PUPP_kat<-factor(data2$mat_PUPP_kat, levels = c("nelaike",
"nepatenkinamas", "patenkinamas", "pagrindinis", "aukstesnysis"))

data2<-data2 %>% mutate(liet_PUPP_kat = ifelse(is.na(data2$lietuviu_PUPP),
"nelaike", ifelse(data2$lietuviu_PUPP ==1 | data2$lietuviu_PUPP ==2|
data2$lietuviu_PUPP ==3,"nepatenkinamas", ifelse(data2$lietuviu_PUPP ==4 |
data2$lietuviu_PUPP ==5, "patenkinamas", ifelse(data2$lietuviu_PUPP ==6 |
data2$lietuviu_PUPP ==7 | data2$lietuviu_PUPP == 8, "pagrindinis",
ifelse(data2$lietuviu_PUPP==9 | data2$lietuviu_PUPP==10, "aukstesnysis",
"x"))))))))
data2$liet_PUPP_kat<-factor(data2$liet_PUPP_kat, levels = c("nelaike",
"nepatenkinamas", "patenkinamas", "pagrindinis", "aukstesnysis"))

#Atskiriama 4000 stebėjimu kryžminiai validacijai (angl.-cross-validation)
set.seed(19970107)
train_ind <- sample(seq_len(nrow(data2)), size = 4000)
data3.test<-data2[train_ind, ]
data3.train<-data2[-train_ind, ]

```

```

#Logistinė regresija 1 modeliui:

mod1=with(data3.train, glm(istojo~pajamos+lytis+gyvenviete+
etnine_mazuma+bent_vienas_aukstos_kval+kvalif_mama+kvalif_tetis+
                        liet_PUPP_vid+mat_PUPP_vid+ egz_suma,
                        family=binomial(link="logit")))

summary(mod1)

round(exp(mod1$coefficients),3) #galimybių santykis
mod1.reduced<-with(data3.train, glm(istojo ~1, family=binomial(link="logit") ))
anova(mod1.reduced, mod1, test="Chisq")


rkvadr1<-1-mod1$deviance/mod1$null.deviance
rkvadr1
library(QuantPsyc)
ClassLog(mod1, data3.train$istojo)


probabilities <-predict(mod1, data3.test, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
table(predicted.classes, data3.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/
sum(table(predicted.classes, data3.test$istojo)))*100


#Logistinė regresija 2 modeliui:

mod2=with(data3.train, glm(istojo~pajamos+lytis+etnine_mazuma+
gyvenviete+bent_vienas_aukstos_kval+kvalif_mama+kvalif_tetis+
                        liet_PUPP_kat+mat_PUPP_kat+ egz_suma,
                        family=binomial(link="logit")))

summary(mod2)

round(exp(mod2$coefficients),3)
mod2.reduced<-with(data3.train, glm(istojo ~1, family=binomial(link="logit") ))
anova(mod2.reduced, mod2, test="Chisq")
rkvadr2<-1-mod2$deviance/mod2$null.deviance

```

```

rkvadr2
ClassLog(mod2, data3.train$istojo)
probabilities <-predict(mod2, data3.test, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
table(predicted.classes, data3.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/
sum(table(predicted.classes, data3.test$istojo)))*100

#Logistinė regresija 3 modeliui :

mod3=with(data3.train, glm(istojo~pajamos+lytis+gyvenvietae+
etnine_mazuma+bent_vienas_aukstos_kval+kvalif_mama+kvalif_tetis+
mat_PUPP_0+liet_PUPP_0+mat_PUPP_laike+liet_PUPP_laike+egz_suma,
family=binomial(link="logit")))
summary(mod3)
round(exp(mod3$coefficients),3)
mod3.reduced<-with(data3.train, glm(istojo ~1, family=binomial(link="logit") ))
anova(mod3.reduced, mod3, test="Chisq")
rkvadr3<-1-mod3$deviance/mod3$null.deviance
rkvadr3
library(QuantPsyc)
ClassLog(mod3, data3.train$istojo)

probabilities <-predict(mod3, data3.test, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
table(predicted.classes, data3.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/
sum(table(predicted.classes, data3.test$istojo)))*100

#MCA/(PCA) komponencijų sukūrimas:
data2<- data2 %>% mutate(gyvenvietae_01 = ifelse(gyvenvietae == "Miestas", 1, 0))
#duomenys, naudojami komponentams sukurti sudedami i viena duomenu masyva:

```

```

pca<-as.data.frame(cbind(data2$pajamos, data2$kvalif_mama, data2$kvalif_tetis,
data2$bent_vienas_aukstos_kval, data2$gyvenvieta_01))

#kintamieji paverčiami faktoriais:
pca$V1<-as.factor(pca$V1)
pca$V2<-as.factor(pca$V2)
pca$V3<-as.factor(pca$V3)
pca$V4<-as.factor(pca$V4)
pca$V5<-as.factor(pca$V5)
#MCA analizeės rezultatai:
res.mca<-MCA(pca, graph = FALSE)
print(res.mca)
#tikrinių reikšmių radimas:
eig.val <- get_eigenvalue(res.mca)
eig.val
#Grafikas:
fviz_screepplot(res.mca, "variance",addlabels = TRUE, ylim = c(0, 20))
fviz_screepplot(res.mca, "eigenvalue",addlabels = TRUE, ylim = c(0, 1),
xlab="Pagrindinė komponentė", ylab="Tikrinė reikšmė (Eigenvalue)")

#reikšmės:
ind <- get_mca_ind(res.mca)
kord<-as.data.frame(ind$coord)
colnames(kord)<-c("dim1", "dim2", "dim3", "dim4", "dim5")
#Sujungiamo gautos komponentės ir duomenis:
pca.data.full<-cbind.data.frame(data2, kord)
#Atskiriame tuos pačius 4000 stebinių testavimu:
pca.data2.test<-pca.data.full[train_ind, ]
pca.data2.train<-pca.data.full[-train_ind, ]

#Logistinė regresija 4 modeliui :

```

```

mod4=with(pca.data2.train, glm(istojo~
                                dim1+dim2+dim3+
                                lytis+etnine_mazuma+
                                liet_PUPP_kat+mat_PUPP_kat+
                                egz_suma,family=binomial(link="logit")))

summary(mod4)
round(exp(mod4$coefficients),3)
mod4.reduced<-with(data3.train, glm(istojo ~1, family=binomial(link="logit") ))
anova(mod4.reduced, mod4, test="Chisq")
rkvadr<-1-mod4$deviance/mod4$null.deviance
rkvadr
ClassLog(mod4, data3.train$istojo)
probabilities <-predict(mod4, pca.data2.test, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
table(predicted.classes,pca.data2.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/
sum(table(predicted.classes, data3.test$istojo)))*100

# MCA su Matematika ir lietuvių kategorijomis
#duomenys, naudojami komponentėms sukurti sudedami į vieną duomenų masyvą:
MCA<-as.data.frame(cbind(data2$mat_PUPP_kat, data2$liet_PUPP_kat))
#kintamuosius pakeičiame į faktorius:
MCA$V1<-as.factor(MCA$V1)
MCA$V2<-as.factor(MCA$V2)
#atliekama MCA analizė
#Tikrines reikšmes:
res.mca<-MCA(MCA, graph = FALSE)
eig.val <- get_eigenvalue(res.mca)
eig.val
#Grafikas:
fviz_screplot(res.mca, addlabels = TRUE, ylim = c(0, 45))
fviz_screplot(res.mca, "eigenvalue",addlabels = TRUE, ylim = c(0, 1),

```

```

xlab="Pagrindinė komponentė", ylab ="Tikrinė reikšmė (Eigenvalue)")
#Gaunamos komponentės
ind <- get_mca_ind(res.mca)
ind

kord2<-as.data.frame(ind$coord)
colnames(kord2)<-c("MCA1", "MCA2", "MCA3", "MCA4", "MCA5")
#Visi duomenis sujungiami į bendrą masyvą:
MCA_PCA_data<-cbind.data.frame(kord, data2, kord2 )
#4000 reikšmiu atskiriama ir paliekama modelio testavimui
MCA_PCA_data.test<-MCA_PCA_data[train_ind, ]
MCA_PCA_data.train<-MCA_PCA_data[-train_ind, ]

#Logistinė regresija 5 modeliui:

mod5=with(MCA_PCA_data.train, glm(istojo~
                                dim1+dim2+dim3+MCA1+
                                lytis+etnine_mazuma+
                                egz_suma, family=binomial(link="logit")))

summary(mod5)
round(exp(mod5$coefficients),3)
mod5.reduced<-with(MCA_PCA_data.train, glm(istojo ~1,
family=binomial(link="logit") ))
anova(mod5.reduced, mod5, test="Chisq")
rkvadr5<-1-mod5$deviance/mod5$null.deviance
rkvadr5

library(QuantPsyc)
ClassLog(mod5, MCA_PCA_data.train$istojo)
probabilities <-predict(mod5, MCA_PCA_data.test, type = "response")
predicted.classes<-ifelse(probabilities >0.5, "1","0")
table(predicted.classes, MCA_PCA_data.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/

```

```

sum(table(predicted.classes, data3.test$istojo))*100

#bandomasis modelis
mod6.test=with(MCA_PCA_data.train, glm(istojo~
                                dim1+dim2+dim3+MCA1+MCA2+
                                lytis+etnine_mazuma+
                                egz_suma, family=binomial(link="logit")))

summary(mod6.test)
round(exp(mod6.test$coefficients),3)
mod6.test.reduced<-with(MCA_PCA_data.train, glm(istojo ~1,
family=binomial(link="logit") ))
anova(mod6.test.reduced, mod6.test, test="Chisq")
rkvadr5<-1-mod6.test$deviance/mod6.test$null.deviance
rkvadr5
library(QuantPsyc)
ClassLog(mod6.test, MCA_PCA_data.train$istojo)
probabilities <-predict(mod6.test, MCA_PCA_data.test, type = "response")
predicted.classes<-ifelse(probabilities >0.5, "1","0")
table(predicted.classes, MCA_PCA_data.test$istojo)
(sum(diag(table(predicted.classes, data3.test$istojo)))/
sum(table(predicted.classes, data3.test$istojo))*100

```