

Studentų nubyrejimo Vilniaus universitete vertinimas taikant mašininio mokymo metodus

Arvydas Norgėla, Jurgita Markevičiūtė

Vilniaus Universitetas Matematikos ir informatikos fakultetas
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius
arvydas.norgela@mif.stud.vu.lt

Santrauka. Straipsnyje apžvelgiama studentų nubyrejimo problema. Aptariamos priežastys, panašūs tyrimai šia tema. Atliekama Vilniaus universiteto turimų duomenų analizė apie individualius studentus. Duomenims pritaikomi logistinės regresijos ir k-artimiausių kaimynų klasifikatoriaus modeliai. Palyginus gautus rezultatus, k-artimiausių kaimynų klasifikatorius tiko geriau – tiksliau klasifikavo visus studentus, bei studentus, kurie metė studijas.

Raktiniai žodžiai: studentų nubyrejimas, logistinė regresija, k-artimiausių kaimynų klasifikatorius, mašininis mokymas, klasifikavimas.

1 Įvadas

Visuose pasaulio universitetuose yra susiduriama su studentų nubyrejimo problema. Pasirinkimas mesti studijas gali kilti dėl labai įvairių priežasčių bei jų kombinacijų, kurias nebūtinai gali kontroliuoti pats universitetas. Tačiau, žinant šias priežastis ir tiriant, kaip jos veikia tam tikro universiteto studentų su studijomis susijusius sprendimus, galima prognozuoti, kurie studentai ketina mesti studijas. Šio tyrimo tikslas yra apžvelgti kituose darbuose naudojamus metodus ir kokiems duomenims jie taikomi; palyginti juos su Vilniaus universiteto turimais duomenimis apie studentus, išsirinkti prasmingus kintamuosius, su jais sudaryti modelius, palyginti jų klasifikavimo tikslumą. Straipsnyje pirmą kartą bandoma pritaikyti klasifikavimo modelius Vilniaus universiteto studentų duomenims.

2 Literatūros apžvalga

Šiame straipsnyje studentų nubyrejimas yra suprantamas kaip išėjimas iš universiteto negavus diplomo [1]. Gali būti labai daug studentų pasitrau-

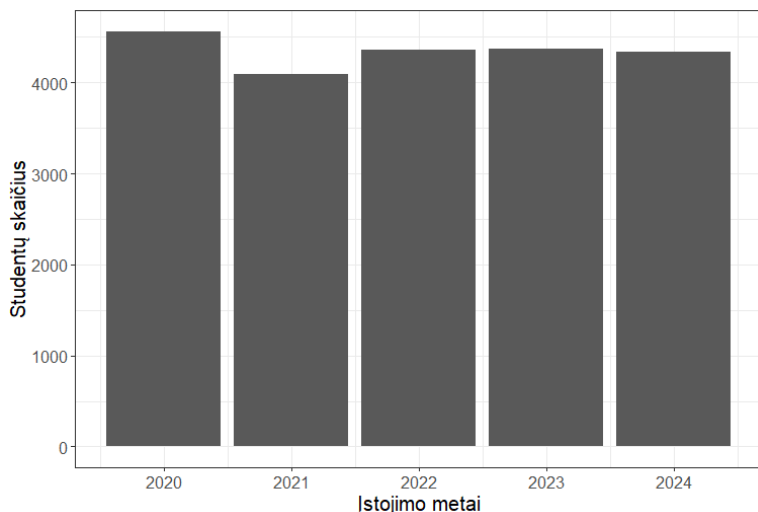
kimo iš studijų priežasčių. Joms aprėpti straipsnyje [1] išskiriamos keturios perspektyvos, per kurias galima žvelgti į studentų nubyrimo priežastis: sociologinė, psichologinė, ekonominė ir fazinė perspektyvos. Sociologinė perspektyva siūlo žvelgti į bendrą studento santykį su universitetu [6]. Psichologinė perspektyva skiria dėmesį studentų psichologinėms savybėms ir kaip jos veikia jų santykį su universitetu [7]. Ekonominė perspektyva remiasi ekonomikos teorija, žvelgia į studentą kaip į racionalų agentą, kuris sugeba įvertinti studijų naudą ir ją palyginti su kaštais [8]. Tarp šių perspektyvų yra persidengimo. Jos visos pabrėžia studento lūkesčius studijoms ir teigia, kad studentas įvertina studijų naudą. Fazinė perspektyva apjungia visas kitas, siūlydama į nubyrimą žvelgti kaip į ilgalaikį procesą, kurio metu stebima individo vidinė būsena ir jo santykis su aplinka [9]. Ji mažiausiai pritaikoma modelių kūrimui, kadangi tokius detalius duomenis gauti būtų sudėtinga. Dažniausiai tyrimuose naudojami su sociologine perspektyva susiję kintamieji, tokie kaip rasė [3], lytis [3], [5], egzaminų pažymys [3], gimimo metai [3], įstojimo metai [3], dėstytojų skaičius universitete [4], tėvų išsilavinimas [5]. Taip pat yra bandoma sieti studentų nubyrimą su ekonominiais rodikliais. Straipsnyje [5] į modelį įtraukiamas nedarbo lygis, infliacija ir bendras vidaus produktas.

Tyrimuose duomenys dažniausiai būna apie individualius studentus. Tai leidžia taikyti labai įvairius metodus. Straipsnyje [3], klasifikuojant studentus į dvi grupes, taikoma logistinė regresija. Šis metodas lyginamas su k-artimiausių kaimynų klasifikatoriumi bei atsitiktinio miško algoritmu. Logistinė regresija klasifikavo studentus tiksliausiai. Straipsnyje [2] panašioms duomenims naudojami mašininio mokymo metodai. Lyginami atsitiktinio miško, tiesinės diskriminantinės analizės ir atraminių vektorių klasifikatoriaus metodai. Atraminių vektorių klasifikatorius buvo tiksliausias, kuris studentus klasifikavo su tikslumu nuo 0,79 iki 0,94, priklausomai nuo mokyklos. Taip pat yra tyrimų, kurie naudoja kitokius duomenis. Pavyzdžiui, [4] straipsnyje naudojami paneliniai duomenys. Tyrime pastebėta, kad yra neigiama koreliacija tarp studentų nubyrimo ir dėstytojų skaičiaus.

3 Duomenų analizė

Mūsų turimi duomenys yra apie individualius studentus. Kaip daugumoje tyrimu, yra kintamieji žymintys lytį, studijų kryptį, bei stojamąjį balą. Apie kai kurių studentų stojamąjį balą nėra duomenų, kas kelia keblumą sudarinė-

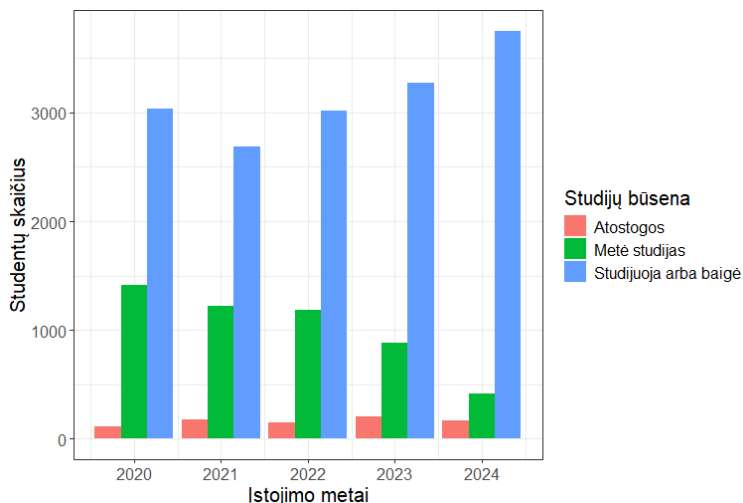
jant modelius su šiuo kintamuoju. Duomenys apima 2020 – 2024 metų laikotarpį. Iš viso duomenyse yra 20231 stebėjimai, stebimi 36 kintamieji, tokie kaip studento lytis, studijų semestras, kursas, įstojimo metai, fakultetas, studijų programa, pakopa, tipas, forma, būseną, priėmimo balas, nesimokymo priežastis, mokėjimo forma. Nuo [2] ir [3] straipsniuose naudotų duomenų mūsų duomenys skiriasi tuo, kad turime duomenis apie studijų semestrą bei mokėjimo formą. Tai mums leidžia pamatyti, kaip studentų sprendimai, susiję su studijų metimu, kinta ilgiau studijuojant bei ar yra skirtumas tarp studentų, kurių studijas apmoka valstybė, ir studentų, kurie už studijas moka patys. Studentų skaičius pagal metus pateikiamas 1 paveikslėlyje.



1 pav. Studentų skaičius pagal metus grafikas, 2020 – 2024 metų laikotarpis

Duomenyse taip pat yra kintamasis, kuris nurodo studento studijų būseną. Yra galimos šios būsenos: sustabdytos studijos, sustabdytos studijos baigėsi, akademinėse atostogose, akademinės atostogos baigėsi, komandiruotėje, baigė studijas, pašalintas, studijuoja ir būseną neapibrėžta. Eilutės su „būseną neapibrėžta“ buvo pašalintos, kadangi jų neįmanoma priskirti jokiai grupei. Taip pat buvo pašalinti visi mainų studentai. Tyrimo tikslas yra klasifikuoti studentus į dvi grupes: tuos, kurie liks ir gaus diplomą, ir tuos, kurie mes studijas. Tad turimos studijų būsenos perkoduotos į teigiamas (studentas dar studijuoja arba gavo diplomą), neigiamas (studentas dėl kaž-

kokios priežasties nebestudijuoja) ir neutralias (baigtiniam laikui sustabdytos studijos) būsenas. Teigiamoms būsenoms priskiriamos šios kategorijos: komandiruotėje, baigė studijas, studijuoja. Neigiamoms: sustabdytų studijų trukmė baigėsi, akademinės studijos baigėsi, pašalintas. Neutralioms: sustabdytos studijos, akademinėse atostogose.

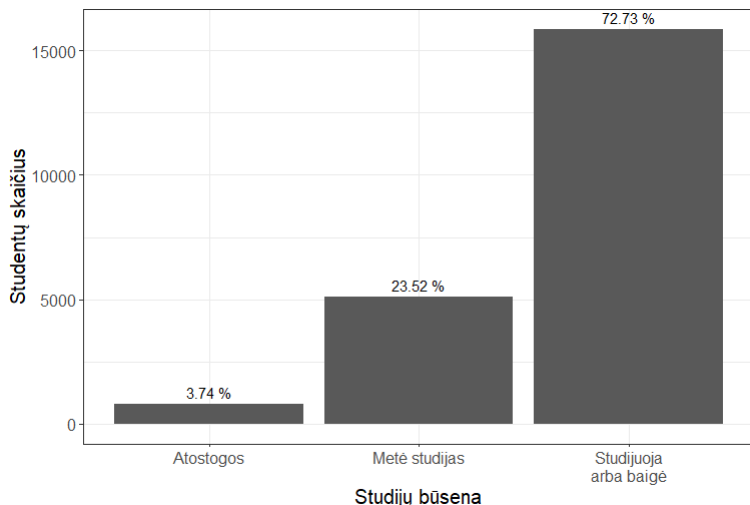


2 pav. Studijų būsenų pagal metus grafikas, 2020 – 2024 metų laikotarpis

2 paveikslėlyje pavaizduota kiek studentų yra kokioje iš anksčiau aptartų būsenų ir kaip šie skaičiai kinta per mūsų tiriamą laikotarpį. Galima pastebėti dvi tendencijas. Pirmoji yra didėjantis studijuojančių arba baigusių studentų skaičius. Antra tendencija yra mažėjantis metusių studijas studentų skaičius.

Duomenyse taip pat galime pastebėti dar vieną problemą – nesubalansuotas imtis. 3 paveikslėlyje pavaizduotame grafike galima pamatyti kiek studentų atsiduria kokioje grupėje. Taip pat parodyta, kokią dalį visų studentų sudaro konkreti grupė. Matome, kad didžioji dalis studentų dar studijuoja arba sėkmingai baigė studijas – apie 72,73 %. Studijas metė 23,52 % studentų, o likusi dalis yra arba akademinėse atostogose, arba laikinai sustabdyusių studijas. Nesubalansuotos imtys gali lemti modelio prisitaikymą priskirti dažniau pasitaikančias kategorijas tiksliau nei kitas, taip iškraipant rezultatus. [5] straipsnyje taip pat pabrėžiama, kad buvo susidurta su ne-

subalansuotų imčių problema. Autoriai rekomendavo naudoti tiek duomenų lygio, tiek algoritmo lygio metodus, kad būtų išvengta šios problemos. Pavyzdžiui, naudoti mašininio mokymo algoritmus, kuriuose yra iš karto įtraukti nesubalansuotų imčių problemos sprendimai.



3 pav. Studentų studijų būsenų pagal metus grafikas, 2020–2024 metų laikotarpis

4 Modeliavimas

Turimiems duomenims galima taikyti kelis modelius. Jei norima klasifikuoti juos į tris kategorijas, galima naudoti daugianarę logistinę regresiją arba k-artimiausių kaimynų klasifikatorių. Kitu atveju, galima pašalinti atostogų būseną iš duomenų ir taikyti metodus, minėtus straipsniuose. Duomenims pritaikius agregavimą, galima taikyti ir laiko eilučių metodus. Kaip siūloma [4] straipsnyje, būtų įmanoma prijungti makroekonominis duomenis. Tai leistų tirti studentų nubybrėjimą iš ekonominės perspektyvos.

Atsižvelgus į apžvelgtus straipsnius [2] ir [3], bei dėl paprastesnės modelių interpretacijos ir tikslesnio klasifikavimo, buvo nuspręsta iš duomenų pašalinti atostogų būseną, taip paliekant dvi grupes. Turimų stebėjimų skaičius sumažėjo iki 19464 studentų. Tuomet buvo pritaikyta logistinė regresija ir k-artimiausių kaimynų klasifikatorius. Modeliai pasirinkti pagal [3] straipsnyje naudotus modelius, kadangi norėta išbandyti, ar tie patys modeliai

tiks mūsų duomenims. Taip pat tai leido mums palyginti gautus rezultatus. Duomenys buvo suskirstyti į mokymo ir testavimo aibes. 80 % duomenų pateko į mokymo aibę, likę 20 % į testavimo. Atrinkti dominantys kintamieji: lytis, studijų semestras, fakultetas, mokėjimo forma ir priėmimo balas. Kintamieji atrinkti remiantis [2] ir [3] straipsniuose naudojamais kintamaisiais. Nuo šių straipsnių pasirinkti kintamieji skiriasi tuo, kad įtraukiama studijų finansavimo forma, bei nenaudojamas amžiaus kintamasis, vietoje jo turime kintamąjį, kuris parodo, kelintame semestre tuo metu buvo studentas. Studentai, kurie priėmimo balo neturėjo, buvo pašalinti. Lytis ir mokėjimo forma buvo perdaryti į binarinius kintamuosius, o fakulteto stulpelis buvo perdarytas į pseudokintamąjį. Kadangi dauguma kintamųjų yra kategoriniai, multikolinearumo problemos duomenyse nebuvo. Pritaikant modelius šiems duomenims buvo atliekamas vienkartinis eksperimentas. Logistinės regresijos atveju, dėl nesubalansuotų imčių buvo išbandytos įvairios slenksčio reikšmės. Geriausi rezultatai buvo pasiekti su 0,6 slenksčiu. Lyties, semestro, mokėjimo formos ir priėmimo balo kintamieji modelyje yra statistiškai reikšmingi. Dauguma fakulteto kintamojo kategorijų buvo statistiškai reikšmingos, tad šį kintamąjį taip pat buvo nuspręsta palikti modelyje. Tai parodo, kad fakulteto pasirinkimas gali paveikti galimybę, kad studentas mes studijas. Koeficientų reikšmės parodė, kad vyrams yra didesnė galimybė mesti studijas, o taip pat ir tiems studentams, kurie patenka į valstybės nefinansuojamas vietas. Galimybė mesti studijas mažėja didėjant semestro reikšmei, tai yra, kuo ilgiau studentas mokosi universitete, tuo mažesnė galimybė, kad jis mes studijas. Tas pats pastebėta ir apie priėmimo balą – didėjant priėmimo balui, mažėja galimybė, kad studentas mes studijas.

Pritaikytam k-artimiausių kaimynų klasifikatoriui buvo išbandytos įvairios k reikšmės. Geriausi rezultatai pasiekti su 5 kaimynais. Modelių klasifikavimo metrikos pateiktos 1 lentelėje. Šioje lentelėje galima pamatyti, kad k-artimiausių kaimynų klasifikatorius visais atžvilgiais, išskyrus jautrumą, mūsų duomenims tiko labiau nei logistinė regresija. Straipsnyje [3], taikant logistinę regresiją buvo pasiektas 0,6659 tikslumas, o su k-artimiausių kaimynų klasifikatoriumi – 0,6460 tikslumas, tai yra, mūsų sukurti modeliai žymiai geriau klasifikuoja studentus. Lyginant su [2] straipsniu, galima pastebėti, kad mūsų sukurti modeliai, ypač k-artimiausių kaimynų, nedaug atsiliko nuo straipsnyje sudaryto atraminių vektorių klasifikatoriaus modelio. Jo tikslumas, priklausomai nuo mokyklos, buvo tarp 0,77 ir 0,94, jautrumas tarp 0,73 ir 0,99 bei specifiškumas tarp 0,76 ir 0,94.

1 lentelė. Modelių metrikos testavimo aibe

Modelis	Tikslumas	Jautrumas	Specifiškumas	F1
Logistinė regresija	0,8439	0,9693	0,4318	0,9049
k-artimiausių kaimynų	0,9147	0,9163	0,9074	0,9462

5 Išvados

Straipsnyje buvo apžvelgta su studentų nubyrėjimu susijusi literatūra. Pastebėta, jog klasifikavimui naudojami įvairūs regresijos ir mašininio mokymo metodai studentų klasifikavimui į studijuojančius ir metusius studijas. Buvo atlikta Vilniaus universiteto turimų duomenų apie studentus analizė. Duomenys apima 2020 – 2024 metų studijų laikotarpį. Studentai buvo suskirstyti į tris grupes: studijuojančius arba baigusius studijas, metusius studijas ir esančius atostogose. Kadangi kitų straipsnių modeliuose naudotos tik dvi kategorijos, buvo nuspręsta neįtraukti į modelius studentų, kurie buvo priskirti atostogaujančių grupei. Tuomet buvo sudaryti logistinės regresijos ir k-artimiausių kaimynų klasifikatoriaus modeliai. Mūsų modeliai nuo kituose straipsniuose taikytų modelių skyrėsi tuo, kad buvo įtraukti studijų semestro ir mokėjimo formos kintamieji. Įvertinti logistinės regresijos koeficientai parodė, jog studentai, patekę į valstybės nefinansuojamas vietas, ir vyrai turi didesnę galimybę mesti studijas. Didėjant semestrų skaičiui, tai yra, ilgiau mokantis universitete, bei didėjant priimamojo balo vertei, galimybė mesti studijas mažėja. Taip pat pastebėta, kad fakulteto pasirinkimas gali turėti poveikį galimybei mesti studijas. Turimiems duomenims k-artimiausių kaimynų klasifikatorius tiko labiau. Modelio tikslumas testavimo aibe buvo apie 0,9147 su 0,9163 jautrumu ir 0,9074 specifiškumu. Šis modelis taip pat tiko mūsų duomenims labiau, kadangi jis ne tik tiksliau bendrai klasifikavo visus studentus, bet ir studentus, kurie metė studijas.

Literatūra

[1] Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: a literature review. *Rev Educ*, 8, 614–652.

[2] Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student Dropout Prediction. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I*, 12163, 129–140.

[3] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting Student Dropout in Higher Education.

- [4] Gitto, L., Minervini, L. F., & Monaco, L. (2016). University dropouts in Italy: Are supply side characteristics part of the problem? *Economic Analysis and Policy*, Volume 49, 108–116.
- [5] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11), 146.
- [6] Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1, 64–85
- [7] Bean, J. P., & Eaton, S. B. (2000). A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the student departure puzzle* (48–61). Nashville: Vanderbilt Univ. Press.
- [8] Becker, R., & Hecken, A. E. (2009). Higher Education or Vocational Training? An Empirical Test of the Rational Action Model of Educational Choices Suggested by Breen and Goldthorpe and Esser. *Acta Sociologica*, 52(1), 25–45.
- [9] Heublein, U. (2014). Student drop-out from German higher education institutions. *European Journal of Education*, 49(4), 497–513.