Požymių eliminavimas studijų nutraukimo prognozavimo uždavinyje

Amanda Balnionytė

Vilniaus Gedimino technikos universitetas, Saulėtekio al. 11, Vilnius, Lietuva
amanda.balnionyte@gmail.com

Santrauka. Šiame straipsnyje tiriamas požymių atrankos ir eliminavimo poveikis studentų studijų nutraukimo prognozavimo kokybei. Duomenų rinkinį sudaro skirtingų grupių požymiai: akademiniai, demografiniai, instituciniai, socioekonominiai, elgsenos ir išvestiniai elgsenos. Atliekama požymių svarbos analizė ir nuoseklus jų mažinimas etapais, siekiant nustatyti optimalų požymių rinkinį. Rezultatai rodo, kad didžiausią įtaką turi elgsenos dinamika ir jos išvestiniai rodikliai, o perteklinių požymių šalinimas supaprastina modelį, tačiau kartu mažina jo gebėjimą tiksliai identifikuoti studijų nutraukimo atvejus.

Raktiniai žodžiai: studijų nutraukimas, požymių atranka, mokymosi analitika, elgsenos duomenys, išvestiniai požymiai.

1 Įvadas

Studentų studijų nutraukimas yra reikšminga aukštojo mokslo problema, turinti akademinių, socialinių ir ekonominių pasekmių. Ankstyvas rizikos studentų identifikavimas leidžia taikyti prevencines priemones, todėl prognozavimo modeliai tampa svarbia tyrimų kryptimi. Pastaruoju metu vis plačiau taikomi mokymosi analitikos (angl. learning analytics) metodai, leidžiantys analizuoti studentų elgseną ir nustatyti su iškritimu susijusius dėsningumus.

Vienas pagrindinių iššūkių yra tinkamas požymių parinkimas. Duomenyse dažnai gausu įvairių kintamųjų, tačiau ne visi jie yra informatyvūs, o pertekliniai požymiai gali bloginti modelio veikimą. Todėl požymių atranka ir eliminavimas yra esminis etapas, leidžiantis išskirti svarbiausius veiksnius.

Šio straipsnio tikslas – įvertinti skirtingų požymių grupių svarbą ir nustatyti optimalų požymių rinkinį studentų iškritimo prognozavimui. Analizuojami įvairių tipų požymiai, ypatingą dėmesį skiriant elgsenos dinamikai, o jų svarba vertinama taikant statistines metrikas ir nuoseklų eliminavimą. Gauti rezultatai leidžia pagerinti modelio efektyvumą ir atskleisti pagrindinius iškritimą lemiančius veiksnius.

DOI: https://doi.org/10.15388/LMITT.2026.1

2 Literatūros analizė

Prognozinių modelių, skirtų studentų studijų nutraukimo rizikai nustatyti, veiksmingumas glaudžiai susijęs su pasirenkamų požymių kokybe ir kiekiu. Įvairūs literatūros šaltiniai nurodo, kad aukštojo mokslo duomenų rinkiniai dažnai pasižymi didele įvairove, apimančia akademinius, demografinius, socioekonominius, elgsenos ir institucinius kintamuosius, todėl neadekvatus požymių atrankos procesas gali lemti modelių persimokymą ir sumažinti jų gebėjimą apibendrinti rezultatus naujuose duomenyse. Atlikta nemažai mokslinių tyrimų, iš kurių matyti, kad požymių atrankos metodai atlieka dvejopą funkciją: mažina skaičiavimo kaštus ir kartu pagerina modelių sprendimų suprantamumą, kas ypač svarbu švietimo kontekste, kur prognoziniai sprendimai turi būti paaiškinami pedagoginiu ir administraciniu lygmeniu.

Vienas dažniausiai taikomų metodologinių požiūrių yra filtravimo metodai, kurie remiasi statistiniu ryšio tarp atskirų požymių ir tikslinio kintamojo vertinimu dar prieš modelio mokymą. Tyrėjai išskiria šį argumentą, kad filtravimo metodai pasižymi dideliu skaičiavimo efektyvumu, nes leidžia greitai įvertinti didelį kintamųjų skaičių nepriklausomai nuo pasirinkto mašininio mokymosi algoritmo. Mokslinėje literatūroje teigiama, kad koreliacijos koeficientai, chi kvadrato testai bei informacinės teorijos metrikos, tokios kaip tarpusavio informacija ar informacijos prieaugis, yra ypač tinkami pirminei požymių atrankai, pašalinant nereikšmingus ar tarpusavyje stipriai susijusius kintamuosius [1].

Moksliniuose šaltiniuose nemažai dėmesio skiriama ir filtravimo metodų ribotumams, ypač tuomet, kai studentų elgseną lemia sudėtingos požymių sąveikos. Tyrėjas taiko skirtingus tyrimo metodus ir parodo, kad vienmačiai statistiniai testai dažnai nepajėgia aptikti sąlyginių priklausomybių, kurios tampa svarbios prognozuojant studijų nutraukimą ankstyvosiose studijų stadijose. Šiame kontekste literatūroje minimi Relief tipo algoritmai, kurie vertina požymių svarbą pagal jų gebėjimą atskirti panašius studentus iš skirtingų klasių, tokiu būdu geriau atspindėdami kompleksinę švietimo duomenų struktūrą [1].

Kita plačiai nagrinėjama metodų grupė yra apgaubiamieji metodai, kurie požymių atranką sieja tiesiogiai su konkretaus prognozinio modelio veikimu. Šie metodai iteratyviai vertina skirtingus požymių poaibius pagal modelio tikslumo pokyčius, todėl leidžia identifikuoti tokias požymių kombinacijas, kurios nėra aptinkamos taikant filtravimo metodus. Moksliniai tyrimai parodė, kad nuoseklios atrankos algoritmai, tokie kaip laipsniškas požymių

įtraukimas ar šalinimas, gali pagerinti prognozavimo rezultatus, tačiau kartu reikalauja žymiai didesnių skaičiavimo resursų [2].

Įterptiniai požymių atrankos metodai literatūroje vertinami kaip kompromisinis sprendimas tarp skaičiavimo efektyvumo ir atrankos tikslumo. Mokslininkai išskiria svarbiausius veiksnius, susijusius su šių metodų populiarumu: požymių svarba vertinama tiesiogiai modelio mokymo metu, todėl nereikia atskiro atrankos etapo. Įterptiniai metodai ypač naudingi, kai siekiama suderinti prognozinį tikslumą su sprendimų paaiškinamumu.

Mokslo tyrimuose vis dažniau pabrėžiama, kad praktikoje efektyviausi yra hibridiniai požymių atrankos sprendimai. Apibendrinant autorių mintis galima spręsti, kad filtravimo metodai dažniausiai taikomi kaip pirmasis etapas, leidžiantis sumažinti dimensiją, o apgaubiamieji ar įterptiniai metodai naudojami galutiniam požymių rinkinio optimizavimui.

3 Metodologija

Tyrime naudojamas apdorotas duomenų rinkinys, sudarytas iš Vilniaus Gedimino technikos universiteto studentų duomenų. Rinkinį sudaro 4036 įrašai su 71 požymiu. Duomenys apima skirtingas požymių grupes: akademinius, istorinius akademinius, demografinius, institucinius, socioekonominius, elgsenos ir išvestinius elgsenos požymius. Akademiniai ir istoriniai akademiniai požymiai aprašo studento pasiekimus, demografiniai – pagrindines charakteristikas, o instituciniai – studijų kontekstą ir administracinę informaciją, įskaitant tikslinį kintamąjį AR_ISKRITO.

Elgsenos požymiai sudaryti iš savaitinių aktyvumo rodiklių (SAVAITE_1-SAVAITE_30), kurie atspindi studento prisijungimų prie Moodle sistemos skaičių. Išvestiniai elgsenos požymiai apibūdina aktyvumo dinamiką, pokyčius ir santykį su grupe. Siekiant struktūruoti laiko sekos duomenis, savaitės sugrupuotos į laikotarpius pagal studijų kalendorių: P1 (1-5 savaitė), P2 (6–10 savaitė), P3 (11-15 savaitė) ir P4 (16-21 savaitė). P4 atitinka sesijos ir pakartotinės sesijos laikotarpį, o savaitės nuo 22 iki 30 priklauso atostogų laikotarpiui. Duomenų rinkinyje naudojami požymiai pateikti 1 lentelėje.

Požymių svarbos vertinimas šiame tyrime grindžiamas kelių tarpusavyje papildančių metodų taikymu. Naudotos trys pagrindinės metrikos: abipusė informacija (angl. mutual information), Pirsono koreliacija ir chi kvadrato testas. Toks derinys leidžia vienu metu įvertinti tiek tiesinius, tiek netiesinius ryšius tarp požymių ir tikslinio kintamojo. Kadangi studentų iškritimas yra kompleksinis reiškinys, vienos metrikos naudojimas galėtų

nepakankamai atspindėti realius ryšius, todėl pasirinktas kombinuotas vertinimo metodas.

Abipusė informacija leidžia nustatyti bendrą priklausomybę tarp požymių ir iškritimo, nepriklausomai nuo ryšio formos. Pirsono koreliacija naudojama įvertinti tiesinį ryšį skaitiniams požymiams, o chi kvadrato testas taikomas kategoriniams požymiams, siekiant nustatyti jų priklausomybę nuo tikslinio kintamojo. Kadangi šios metrikos yra skirtingų skalių, jos buvo normalizuotos min-max metodu, o kiekvienam požymiui apskaičiuotas bendras balas kaip jų suma. Tai leidžia išskirti požymius, kurie yra nuosekliai svarbūs pagal kelis kriterijus, ir sumažina pavienių metrikų šališkumą.

1 lentelė. Duomenų rinkinio sandara.

Požymių grupė	Požymio pavadinimas	
Akademiniai	SEM_VIDURKIS	
Istoriniai akademiniai	STOJ_BALAS	
Demografiniai	AMZIUS	LYTIS
Instituciniai	STUD_KODAS	AR_ISKRITO
	GRUPE	AR_MOBILUMAS
	PROGRAMA	SEMESTRAS
	KALBA	METAI
Socioekonominiai	AR_FINANSAVIMAS	AR_STIPENDIJA
Elgsenos	SAVAITE_1	SAVAITE_16
	SAVAITE_2	SAVAITE_17
	SAVAITE_3	SAVAITE_18
	SAVAITE_4	SAVAITE_19
	SAVAITE_5	SAVAITE_20
	SAVAITE_6	SAVAITE_21
	SAVAITE_7	SAVAITE_22
	SAVAITE_8	SAVAITE_23
	SAVAITE_9	SAVAITE_24
	SAVAITE_10	SAVAITE_25
	SAVAITE_11	SAVAITE_26
	SAVAITE_12	SAVAITE_27
	SAVAITE_13	SAVAITE_28
	SAVAITE_14	SAVAITE_29
	SAVAITE_15	SAVAITE_30

Požymių grupė	Požymio pavadinimas	
Išvestiniai elgsenos	AKT_P1_VID	SANT_KRIT_P1_P2
	AKT_P1_STD	SANT_KRIT_P2_P3
	AKT_P2_VID	SANT_KRIT_P3_P4
	AKT_P2_STD	MAKS_SANT_KRIT
	AKT_P3_VID	MAKS_ABS_KRIT
	AKT_P3_STD	GRUP_AKT_P1_VID
	AKT_P4_VID	GRUP_AKT_P2_VID
	AKT_P4_STD	GRUP_AKT_P3_VID
	POKYTIS_P1_P2	GRUP_AKT_P4_VID
	POKYTIS_P2_P3	GRUP_SKIRT_P1
	POKYTIS_P3_P4	GRUP_SKIRT_P2
	NUL_SAV_P1	GRUP_SKIRT_P3
	NUL_SAV_P2	GRUP_SKIRT_P4

Bazinis modelis šiame tyrime buvo logistinės regresijos klasifikatorius. Modelis realizuotas naudojant apdorojimo seką (angl. *pipeline*), kurioje sujungtas duomenų paruošimas ir modeliavimas. Kategoriniai požymiai buvo transformuoti taikant vienkarštį kodavimą (angl. *one-hot encoding*). Duomenys suskirstyti į mokymo ir testavimo aibes santykiu 80 ir 20, taikant stratifikaciją pagal tikslinį kintamąjį AR_ISKRITO. Modelio stabilumui vertinti naudota kryžminė patikra (angl. *cross-validation*) su 5 dalimis. Atsitiktinumo kontrolei taikyta sėkla (angl. *random state*) lygi 42. Dėl klasių disbalanso naudoti subalansuoti klasių svoriai. Modelio parametrai nustatyti: reguliarizacijos stiprumas C lygus 1, naudojamas optimizavimo algoritmas (angl. *solver*) liblinear, maksimalus iteracijų skaičius lygus 1000.

Tyrime požymių eliminavimas vykdomas etapiniu būdu, nuosekliai mažinant požymių skaičių ir vertinant modelio veikimo pokyčius. Pradiniame etape naudojamas pilnas požymių rinkinys, o vėlesniuose etapuose požymiai šalinami grupėmis pagal jų tipą ir svarbą. Kiekviename etape modelis treniruojamas iš naujo, o jo veikimas vertinamas naudojant klasifikavimo rodiklius, siekiant nustatyti optimalų požymių rinkinį.

Modelio veikimas bus vertinamas naudojant kelias klasifikavimo metrikas: bendrą klasifikavimo tikslumą (angl. *accuracy*), tikslumą (angl. *precision*), jautrumą (angl. *recall*), F1 rodiklį, AUC-ROC kreivės plotą ir klaidų matricą (angl. *confusion matrix*). Bendras tikslumas parodo teisingai klasifikuotų atvejų dalį, tačiau dėl duomenų disbalanso ši metrika nėra pakankama verti-

nimui. Tikslumas įvertina, kiek iš prognozuotų teigiamų atvejų yra teisingi, o jautrumas parodo, kokią dalį realių iškritimo atvejų modelis sugeba aptikti. F1 rodiklis apjungia tikslumą ir jautrumą į vieną balansinį įvertį, todėl leidžia įvertinti bendrą modelio veikimą esant nesubalansuotiems duomenims. AUC metrika parodo modelio gebėjimą atskirti klases nepriklausomai nuo pasirinkto slenksčio. Šiame tyrime pagrindinis dėmesys skiriamas jautrumui ir F1 rodikliui, nes svarbiausia yra kuo tiksliau identifikuoti iškritimo atvejus ir išlaikyti balansą tarp klaidingų teigiamų ir klaidingų neigiamų prognozių.

4 Rezultatai

Tyrime buvo atlikta požymių svarbos analizė, kurios rezultatai buvo pateikti 2 lentelėje. Elgsenos požymių grupėje ryškiausi buvo vidurinio ir vėlyvo semestro laikotarpiai. SAVAITE_14–SAVAITE_19 pasižymėjo stipresniu ryšiu su iškritimu nei ankstyvos savaitės. SAVAITE_1–SAVAITE_5 turėjo mažą svarbą, todėl pradinė elgsena nebuvo pakankamai informatyvi. Tai rodė, kad studentų aktyvumas stabilizavosi vėlesniu laikotarpiu ir tik tuomet tapo patikimu indikatoriumi.

Išvestiniai elgsenos požymiai dominavo visame reitinge. Ypač svarbūs buvo STUD_GRUP_SKIRT_P4, MAKS_SANT_KRIT ir POKYTIS_P2_P3, kurie atspindėjo aktyvumo pokyčius ir atsilikimą nuo grupės. Tai patvirtino, kad svarbiausias veiksnys buvo ne absoliutus aktyvumas, o jo dinamika. Grupiniai rodikliai turėjo vidutinę svarbą, tačiau kartu su individualiais požymiais suteikė papildomos informacijos.

Akademiniai požymiai pasižymėjo vidutiniu reikšmingumu. SEM_VIDURKIS rodė netiesinį ryšį su iškritimu, o STOJ_BALAS turėjo mažesnę įtaką. Instituciniai požymiai buvo nevienareikšmiai: GRUPE buvo svarbi, tačiau kiti, tokie kaip PROGRAMA ar KALBA, turėjo mažą reikšmę. STUD_KODAS buvo laikytinas techniniu požymiu. Socioekonominiai ir demografiniai požymiai turėjo mažiausią įtaką.

Vertinant galimus netikslumus, buvo nustatyta, kad AR_GRIZO pasižymėjo itin stipriu ryšiu su tiksliniu kintamuoju, todėl galėjo kelti duomenų nutekėjimo riziką. Tačiau taip pat buvo nustatyta, kad šis ryšys nebuvo absoliutus, todėl nuspręsta šį požymį palikti.

Tyrime buvo atlikta iteracinė požymių eliminavimo analizė, kurios tikslas buvo sumažinti modelio sudėtingumą išlaikant aukštą prognozavimo kokybę. Pradinėje bazinio modelio konfigūracijoje buvo naudojama 70 požymių.

2 lentelė. Požymių svarbos studijų nutraukimo prognozavime rezultatai.

#	Požymis	Abi	Kor	Chi	Bendras	#	Požymis	Abi	Kor	Chi	Bendras
1	GRUPE	0,72	0,00	1,00	**1,72**	36	AKT_P1_STD	0,24	0,07	0,00	**0,32**
2	GRUP_SKIRT_P4	1,00	0,60	0,00	**1,60**	37	SAVAITE_6	0,14	0,17	0,00	**0,31**
3	MAKS_SANT_KRIT	0,82	0,62	0,00	**1,44**	38	SAVAITE_22	0,07	0,24	0,00	**0,31**
4	AR_GRIZO	0,32	1,00	0,00	**1,32**	39	AR_FINANSAVIMAS	0,12	0,17	0,00	**0,29**
5	SEMESTRAS	0,56	0,60	0,00	**1,16**	40	SAVAITE_25	0,18	0,11	0,00	**0,29**
6	GRUP_AKT_P4_VID	0,76	0,23	0,00	**1,00**	41	NUL_SAV_P2	0,00	0,27	0,00	**0,27**
7	GRUP_SKIRT_P3	0,45	0,52	0,00	**0,98**	42	SAVAITE_28	0,10	0,17	0,00	**0,27**
8	AKT_P4_STD	0,44	0,53	0,00	**0,96**	43	SAVAITE_8	0,08	0,17	0,00	**0,26**
9	GRUP_AKT_P1_VID	0,57	0,37	0,00	**0,94**	44	SANT_KRIT_P2_P3	0,19	0,07	0,00	**0,26**
10	GRUP_AKT_P2_VID	0,60	0,34	0,00	**0,93**	45	SAVAITE_9	0,00	0,24	0,00	**0,24**
11	AKT_P4_VID	0,43	0,50	0,00	**0,93**	46	SANT_KRIT_P1_P2	0,15	0,07	0,00	**0,23**
12	SAVAITE_16	0,40	0,44	0,00	**0,84**	47	SAVAITE_11	0,00	0,23	0,00	**0,23**
13	GRUP_AKT_P3_VID	0,52	0,30	0,00	**0,82**	48	GRUP_SKIRT_P1	0,02	0,17	0,00	**0,19**
14	SEM_VIDURKIS	0,69	0,08	0,00	**0,78**	49	AKT_P2_STD	0,11	0,08	0,00	**0,19**
15	SAVAITE_18	0,37	0,40	0,00	**0,77**	50	AKT_P2_VID	0,00	0,17	0,00	**0,17**
16	SAVAITE_15	0,29	0,42	0,00	**0,71**	51	NUL_SAV_P1	0,11	0,05	0,00	**0,17**
17	SAVAITE_17	0,27	0,37	0,00	**0,64**	52	SAVAITE_27	0,00	0,15	0,00	**0,15**
18	SAVAITE_19	0,28	0,35	0,00	**0,64**	53	AKT_P1_VID	0,14	0,02	0,00	**0,15**
19	GRUP_SKIRT_P2	0,26	0,36	0,00	**0,63**	54	SAVAITE_26	0,00	0,12	0,00	**0,12**
20	SAVAITE_14	0,19	0,41	0,00	**0,60**	55	SAVAITE_30	0,11	0,01	0,00	**0,12**
21	POKYTIS_P2_P3	0,12	0,45	0,00	**0,56**	56	SAVAITE_24	0,07	0,04	0,00	**0,11**
22	AR_STIPENDIJA	0,18	0,33	0,00	**0,51**	57	AMZIUS	0,00	0,11	0,00	**0,11**
23	SANT_KRIT_P3_P4	0,45	0,06	0,00	**0,51**	58	SAVAITE_10	0,00	0,11	0,00	**0,11**
24	STUD_KODAS	0,38	0,13	0,00	**0,51**	59	SAVAITE_7	0,00	0,11	0,00	**0,11**
25	POKYTIS_P1_P2	0,05	0,39	0,00	**0,44**	60	SAVAITE_4	0,00	0,10	0,00	**0,10**
26	SAVAITE_21	0,10	0,33	0,00	**0,43**	61	MAKS_ABS_KRIT	0,00	0,09	0,00	**0,09**
27	SAVAITE_20	0,06	0,37	0,00	**0,43**	62	METAI	0,05	0,00	0,03	**0,08**
28	SAVAITE_29	0,25	0,17	0,00	**0,42**	63	AR_MOBILUMAS	0,00	0,07	0,00	**0,07**
29	SAVAITE_13	0,06	0,36	0,00	**0,42**	64	SAVAITE_3	0,00	0,05	0,00	**0,05**
30	AKT_P3_STD	0,14	0,27	0,00	**0,42**	65	SAVAITE_5	0,00	0,05	0,00	**0,05**
31	AKT_P3_VID	0,04	0,37	0,00	**0,41**	66	SAVAITE_2	0,00	0,04	0,00	**0,04**
32	POKYTIS_P3_P4	0,23	0,18	0,00	**0,41**	67	SAVAITE_1	0,01	0,01	0,00	**0,03**
33	STOJ_BALAS	0,10	0,29	0,00	**0,39**	68	LYTIS	0,00	0,00	0,02	**0,02**
34	SAVAITE_12	0,13	0,25	0,00	**0,37**	69	KALBA	0,00	0,00	0,00	**0,00**
35	PROGRAMA	0,02	0,00	0,30	**0,33**	70	SAVAITE_23	0,00	0,00	0,00	**0,00**

Nors modelis rodė gerus rezultatus, buvo pastebėtas disbalansas tarp tikslumo ir jautrumo, rodantis, kad modelio gebėjimas identifikuoti teigiamus atvejus buvo ribotas. Tolimesniuose etapuose požymių skaičius buvo nuosekliai mažinamas, vertinant modelio veikimo rodiklių pokyčius. Eliminuoti požymiai buvo pateikti 3 lentelėje, o eliminavimo rezultatai – 4 lentelėje.

3 lentelė. Eliminuojami požymiai.

1 etapas	2 etapas	3 etapas	4 etapas	5 etapas
AR_STIPENDIJA	METAI	SAVAITE_1	AR_FINANSAVIMAS	SAVAITE_6
SEM_VIDURKIS	AR_MOBILUMAS	SAVAITE_2	NUL_SAV_P1	SAVAITE_7
STUD_KODAS	AMZIUS	SAVAITE_3	STUD_GRUP_SKIRT_P1	SAVAITE_8
SAVAITE_22	LYTIS	SAVAITE_4	AKT_P1_VID	MAKS_ABS_KRIT
SAVAITE_23	KALBA	SAVAITE_5		SANT_KRIT_P1_P2
SAVAITE_24				SANT_KRIT_P2_P3
SAVAITE_25				AKT_P2_VID
SAVAITE_26				AKT_P2_STD
SAVAITE_27				
SAVAITE_28				
SAVAITE_29				
SAVAITE_30				

4 lentelė. Požymių eliminavimo rezultatai.

Atvejai	Požymių skaičius	Bendras klasifikavimo tikslumas	Tikslumas	Jautrumas	F1 rodiklis	AUC	TN	FP	FN	TP
Bazinis	70	0,94	0,14	0,73	0,29	0,93	748	48	3	8
1 etapas	58	0,93	0,13	0,73	0,23	0,9	744	52	3	8
2 etapas	53	0,93	0,12	0,64	0,2	0,92	745	51	4	7
3 etapas	48	0,93	0,11	0,64	0,19	0,93	741	55	4	7
4 etapas	44	0,93	0,11	0,64	0,19	0,93	741	55	4	7
5 etapas	36	0,92	0,11	0,64	0,18	0,91	737	59	4	7

Pirmiausia buvo nuspręsta pašalinti požymius, kurie buvo netinkami arba per vėlyvi studijų nutraukimo prognozei. Kintamasis STUD_KODAS buvo pašalintas, nes jis buvo skirtas identifikacijai, o ne prognozavimui. Taip pat buvo eliminuoti AR_STIPENDIJA, SEM_VIDURKIS ir elgsenos rodikliai, to-

kie kaip SAVAITE_22–SAVAITE_30, kadangi jie buvo prieinami tik vėlesniu laikotarpiu ir turėjo ribotą prognozinę vertę. Pirmojo eliminavimo etapo metu pašalinus 12 požymių, modelio veikimas nežymiai suprastėjo dėl padidėjusio klaidingai teigiamų atvejų skaičiaus. Nors bendras klasifikavimo tikslumas išliko stabilus, F1 rodiklis sumažėjo.

Antrajame etape buvo pašalinti instituciniai ir demografiniai požymiai, kurie turėjo labai mažą prognozinę vertę. Perėjimas nuo pirmojo prie antrojo eliminavimo etapo rodo, kad modelio gebėjimas identifikuoti teigiamus atvejus pablogėjo. Tuo pačiu nežymiai sumažėjo klaidingai teigiamų prognozių skaičius, o AUC rodiklis padidėjo, rodydamas, kad bendras klasių atskyrimas netgi šiek tiek pagerėjo. F1 rodiklis sumažėjo nedaug, todėl bendras modelio balansas iš esmės išliko panašus, tačiau aiškiai matomas jautrumo sumažėjimas rodo, kad pašalinti požymiai turėjo reikšmės būtent teigiamų atvejų identifikavimui.

Trečiajame etape buvo pašalinti elgsenos požymiai, turėję mažiausią reikšmingumą. Eliminavus SAVAITE_1–SAVAITE_5 požymių skaičius sumažėjo iki 48. Modelio veikimo rodikliai nežymiai suprastėjo dėl padidėjusio klaidingai teigiamų atvejų skaičiaus.

Ketvirtajame etape buvo nuspręsta pašalinti požymius, kurie nė vienoje iš požymių svarbos metrikų nepasiekė 0,2 reikšmės. Tokių požymių duomenų rinkinyje buvo 13. Siekiant detalesnės analizės, buvo nuspręsta eliminuoti tik dalį jų. Buvo pašalintas AR_FINANSAVIMAS bei išvestiniai elgsenos požymiai, susiję su ankstyvuoju pirmuoju laikotarpiu. Sumažinus požymių skaičių modelio rodikliai nepakito, kas rodė, kad pašalinti požymiai neturėjo papildomos prognozinės vertės.

Penktajame etape požymių skaičius buvo sumažintas iki 36, papildomai eliminavus išvestinius elgsenos požymius, kurie nė vienoje iš požymių svarbos metrikų nepasiekė 0,2 reikšmės ir buvo susiję su antruoju laikotarpiu. Modelio kokybė reikšmingai nepasikeitė, tačiau padidėjo klaidingai teigiamų atvejų skaičius, kas rodė nedidelį modelio tikslumo sumažėjimą neigiamų atvejų klasifikavimo atžvilgiu.

5 Išvados

Tyrimas parodė, kad didžiausią įtaką studijų nutraukimo prognozavimui turi elgsenos ir išvestiniai elgsenos požymiai. Ypač svarbūs yra aktyvumo dinamiką apibūdinantys rodikliai, tokie kaip STUD_GRUP_SKIRT_P4, MAKS_SANT_KRIT ir POKYTIS_P2_P3, patvirtinantys, kad lemiamas veiksnys yra ne

aktyvumo lygis, o jo pokyčiai. Taip pat nustatyta, kad vidurinio ir vėlyvo semestro savaitės yra informatyvesnės nei ankstyvosios.

Remiantis požymių eliminavimo rezultatais, racionaliausia požymių atrankos riba laikytinas pirmasis eliminavimo etapas. Šiame etape pašalinus dalį mažiau informatyvių ir prognozavimui netinkamų požymių, modelio veikimas iš esmės išliko stabilus, tačiau jau buvo stebimi pirmieji gebėjimo identifikuoti studijų nutraukimo atvejus silpnėjimo požymiai. Tolimesnis požymių mažinimas šią tendenciją sustiprino – modelis vis rečiau teisingai identifikavo studijų nutraukimo atvejus ir dažniau klydo tiek juos praleisdamas, tiek klaidingai priskirdamas, o tai mažina modelio patikimumą ir blogina jo praktinį pritaikomumą.

Literatūra

[1] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., & Abid, M. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access, 9, 7519–7539.

[2] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146.