Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Ken-Erik Aus

Loodus- ja täppisteaduste valdkonna üliõpilaste väljalangevus

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: MSc Mare Vähi

Tartu 2026

LOODUS- JA TÄPPISTEADUSTE VALDKONNA ÜLIÕPILASTE VÄLJALANGEVUS

Bakalaureusetöö

Ken-Erik Aus

Lühikokkuvõte

Loodus- ja täppisteaduste (LT) erialadel on probleem suure väljalangevusega. Valdkonna spetsialistide järelkasvu tagamiseks on oluline mõista, mis põhjustab õpingute katkestamist. Käesoleva töö eesmärk oli tuvastada olulised mõjutegurid väljalangevusele ja erialavahetusele Tartu Ülikooli LT valdkonna viiel õppekaval. Uuriti valimit 1167 õppeteekonnast 1139 tudengilt, kes immatrikuleeriti aastatel 2019-2023. Õpingute lõppseise prognoositi multinomiaalse ja binaarse logistilise regressiooniga, võrreldes tunnuste valikut tõepärasuhte testi ja Akaike informatsioonikriteeriumi alusel. Väljalangevuse oluliste mõjuteguritena tuvastati esimesel õppeaastal positiivsele tulemusele läbitud ainepunktide määr ja kõikide õpitud semestrite aritmeetiline keskmine hinne. Binaarse mudeli ennustusvõime oli rahuldav ($AUC = 0.903$ testhulgal). Erialavahetuse ennustamine ei õnnestunud: mudeli AUC oli vaid 0.675 ning testhulga 27 erialavahetajast prognoositi õigesti 6. Mõlemas mudelis valisid tõepärasuhte test ja AIC samad tunnused, seega statistiliselt mitteoluliste tunnuste lisamine ennustusvõimet ei parandanud.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Õpingute katkestamine, erialavahetus, multinomiaalne logistiline regressioon.

STUDENT ATTRITION IN THE FACULTY OF SCIENCE AND TECHNOLOGY

Bachelor thesis

Ken-Erik Aus

Abstract

Student attrition is a significant problem in the fields of science and technology. To ensure that the number of specialists in these fields grow, it is important to understand what causes students to leave their studies. The aim of this thesis was to identify significant predictors of dropout and switching programme at the University of Tartu's Faculty of Science and Technology. The sample consisted of 1167 study trajectories from 1139 students on five programmes who enrolled on the school years starting on 2019-2023. Outcomes were predicted using multinomial and binary logistic regression, comparing variable selection by likelihood-ratio test and the Akaike information criterion. As significant predictors of dropout, rate of successfully completed course points in the first academic year and arithmetic mean grade across all studied semesters were identified. The predictive power of the binary model was satisfactory ($AUC = 0.903$ on test set). Prediction of programme switching was not successful: the model's AUC was only 0.675 and of the 27 switchers in the test set, only 6 were predicted correctly. In both models, the likelihood-ratio test and AIC selected identical variable sets, indicating that including statistically non-significant predictors did not improve predictive performance.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Student attrition, programme switching, multinomial logistic regression.

Sisukord

Sissejuhatus

Loodus- ja täppisteaduste õppekavadel on märgatud teistest valdkondadest kõrgemat väljalangevust. Nende erialade ekspertide järelkasv tuleb tagada, sest keerulised ja tehnilised teadmised, mida nendel erialadel omandatakse, aitavad arendada uusi tehnoloogiaid ning laiendada inimeste teadmisi maailmast. Et leevendada väljalangevuse probleemi, on vaja mõista, mis põhjustab õpingute katkestamist loodus- ja täppisteadustes.

Selle uuringu eesmärk oli tuvastada, miks tudengid katkestavad oma õpingud viiel Tartu Ülikooli loodus- ja täppisteaduste valdkonna õppekaval. Erijuhtumina käsitleti ka seda, kui tudeng vahetas eriala, jättes algsele erialale ühe võimaliku spetsialisti vähem. Sõnastati kaks uurimisküsimust.

1. Mis on olulised mõjutegurid väljalangevusele ja erialavahetusele?
2. Kas statistiliselt mitteoluliste tunnuste kasutamine väljalangevuse ja erialavahetuse prognoosimisel parandab ennustusvõimet?

Uuriti valimit Tartu Ülikooli tudengitest, kes immatrikuleeriti õppeaastatel algusega 2019-2023 TÜ loodus- ja täppisteaduste valdkonnas ühel viiest uuritavast erialast. Andmestik koostati 2023/2024 kevadsemestri seisuga ning kaasati ka tudengid, kelle õpingud veel lõppenud ei olnud.

Esimeses peatükis antakse empiiriline taust uuritavale probleemile ning võetakse kokku eelnevalt sarnastes uuringutes tuvastatud mõjutegurid. Teine peatükk kirjeldab metoodikat, mida selles töös uurimisküsimustele vastamiseks kasutatakse. Kolmandas peatükis esitatakse analüüsi tulemused ja arutelu.

1 Taust

Loodus- ja täppisteadused (LT) on Eesti majandus- ja tööturupoliitikas tähtis valdkond. See kuulub laiemasse loodus-, täppis ja tehnikateaduste (LTT) valdkonda (tuntud rohkem inglise keelse lühendiga STEM), mille spetsialistidele on Eesti tööturul suur nõudlus. Viimase kümne aasta jooksul on aga LT erialade populaarsus kahanenud. Kokku lõpetasid STEM aladel hariduse 2020/2021 õppeaastal 28% Eesti ülikoolide tudengitest, mis on 4 protsendipunkti kõrgem kui 2010/2011 aastal. LT valdkonna lõpetajate osakaal oli aga selle aja jooksul kahanenud seitsmelt protsendilt viiele. (Kreegipuu ja Jaggo, 2017; Haridus- ja Teadusministeerium, 2022)

Suur probleem Eesti kõrghariduses on väljalangevus. Kõrghariduse katkestajate määr oli 2020/2021 õppeaastal 13.3 protsenti, Euroopa keskmine oli samal ajaperioodil umbes 10%. Märgatav osa tudengeid katkestavad õpingud esimesel kursusel ning see mure esineb kõige rohkem STEM valdkondades. Loodus- ja täppisteaduste tudengitest ligi 17% katkestasid õpingud 2020/2021 õppeaastal. Hariduse valdkonnas oli väljalangevuse määr vaid 8 protsenti. (Haridus- ja Teadusministeerium, 2022; Eurostat, 2023)

Sarnased probleemid esinevad ka mujal. Ameerika STEM tudengitest 40-50% vahetavad eriala või katkestavad oma hariduse. Ka neil on suur väljalangevus just esimesel kursusel. OECD 2022. aasta raporti andmetel lõpetasid vaid 68% tudengitest STEM valdkonna programmi, terviseteadustes oli see osakaal 80%. Need osakaalud võtsid arvesse ka neid tudengeid, kellel kulus oma eriala lõpetamiseks nominaalajast kuni 3 aastat pikem periood.(Prosper, 2024; OECD, 2022)

1.1 Väljalangevuse mõjutegurid

Oluline väljalangevuse põhjustaja on tudengi õppeedukus ülikooli ajal. Ameerikas tehtud uuringu järgi olid madala keskmise hinde ning suure arvu läbikukutud STEM ainepunktidega tudengid tõenäolisemad õpinguid katkestama (Chen, 2013).

Samuti põhjustab väljalangevust see, kui tulemused loodusteaduste ainetes on madalamad, kui teiste valdkondade ainetes. Läbitud ainete arvu põhjuslikku seost esimese kursuse väljalangevusega kinnitas ka Portugalis tehtud uuringus leitud mudel (Casanova *et al.*, 2023).

Mõjutegurid võivad olla ka ülikoolivälised. Miskolci ülikoolis Ungaris leiavad Varga, Fodor ja Szilágyi (2023) analüüsi tulemusena, et kõige vähem katkestasid õpinguid tudengid, kelle keskmine hinne eelmises haridusastmes oli kõrge. Selles uuringus ei tuvastatud eksamitulemuste otsest olulist mõju väljalangevusele, kuid kõrge ülikoolieelse matemaatikaeksami tulemustega tudengid täitsid rohkem ainepunkte esimesel kahel semestril. Belser *et al.* (2018) kaasasid oma väljalangevust ennustavasse mudelisse matemaatikaeksami tulemused, kuid mõju ei olnud statistiliselt oluline ning tunnus kaasati vaid ennustusvõime parandamiseks.

Avastatud on ka soolisi erinevusi väljalangevuses ning õppeedukuses. Euroopa Liidus on meestudengite väljalangevuse osakaal märgatavalt kõrgem, kui naistudengitel. 2023. aastal katkestasid õpingud 11.3% meestest, kuid naistest vaid 7.7 protsenti (Eurostat, 2024). See trend on aga vastupidine STEM valdkondades. Casanova *et al.* (2023) koostatud mudelis osutus sugu oluliseks väljalangevuse ennustajaks ning meestudengite tõenäosus õpinguid katkestada oli madalam. Belser *et al.* (2018) ei kaasanud sootunnust väljalangevust ennustavasse mudelisse, kuid tõid välja, et tulemust võis mõjutada valimi disain.

Akadeemilise puhkuse võtmine tõstab oluliselt riski, et tudeng katkestab õpingud. Austraalias avastati, et puhkuse võtjatest vähem kui kolmandik olid veel õppimas uuringu ajal või lõpetanud eriala. Nende seas, kes jätkasid ilma peatamiseta õpinguid, oli see osakaal 70%. Need andmed ei keskendunud aga STEM valdkonnale ning kitsama valimiga uuringuid ei leitud. Siiski tuleb kaaluda akadeemilise puhkuse võtmist kui võimalikku mõjutegurit LT valdkonna väljalangevusele. (Naylor, Cox ja Cakitaki, 2023)

1.2 Eriala vahetamine

Loodus- ja täppisteaduste valdkonna lõpetajate arv võib lisaks väljalangevusele ka väheneda eriala vahetamise tõttu. Chen (2013) tuvastas, et suurim mõju eriala vahetamisele on sellel, kui tudeng võtab endale madalama õppekoormuse esimesel kursusel. Madalate tulemustega ning lõpetamata või läbikukutud ainetega tudengid olid samuti tõenäolisemad eriala vahetama. (Chen, 2013)

Peamine erinevus eriala vahetajate ning õpingute katkestajate seas on õppeedukus. Ameerika tudengid, kelle keskmine hinne oli madalam kui 3.0, olid tõenäolisemad ülikoolist välja langema kui 3.5-st kõrgema keskmisega tudengid. Parema sooritusega tudengid olid aga tõenäolisemad vahetama eriala. (Chen, 2013)

2 Metoodika

2.1 Andmete kogumine

Väljalangevuse uurimiseks on efektiivne kasutada ülikooli andmebaasi väljavõtteid. Sellesse meetodisse süvenesid Aulck *et al.* (2017), kes kogusid andmeid Washingtoni ülikooli andmebaasist. Tunnuste seas olid demograafilise tausta indikaatorid, kõrgharidusele eelnevad õppetulemused ja ülikooliaegne õppeedukus. See meetod aga ei võimaldanud uurida tudengi motivatsiooni ja teisi isiklikuma taustaga mõjutajaid. Pedersen ja Nielsen (2024) uuringus kasutati küsimustikku, et pärida LTT tudengite motivatsiooni, hinnangut enda oskustele ja põhjuseid eriala valikuks. Töös keskenduti ka tudengi omaalgatuse hinnangule õppetöös, mis lõpuks ei osutunud oluliseks mõjuteguriks.

Kaitstud andmed nagu sugu ja vanus on sageli kaasatud sarnastes uuringutes. Yu, Lee ja Kizilcec (2021) leidsid, et nende kaasamisel mudelisse ei olnud aga olulist efekti. Kaitstud andmetest on käesolevas töös kasutusel vaid sugu, kuid teiste isikuandmete välja jätmine ei muuda eelneva kirjanduse põhjal mudeli kvaliteeti oluliselt.

2.2 Mudeli koostamise teooria

Logistiline regressioon on enim kasutatud mudel sarnastes uuringutes ja on leitud, et see on teistest efektiivsem (Aulck *et al.*, 2017). Logistilise regressiooni mudel on binaarse sõltuva tunnuse korral järgmisel kujul:

$$\ln\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = \boldsymbol{\beta} \cdot \mathbf{X}_i,$$

kus $\boldsymbol{\beta}$ on mudeli parameetrite vektor ja $\mathbf{X}_i$ on vaatluse i andmete maatriks. Kuna selles töös oli ennustatavaid kategooriaid kolm, ei saanud kasutada aga tavalist logistilist regressiooni.

Multinomiaalne regressioon hindab relatiivsete šansside naturaallogaritmi võrreldes etteantud baaskategooriaga. Matemaatiliselt on mudel kolme kategooriaga järgmisel kujul:

$$\ln\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i,$$
$$\ln\left(\frac{P(y_i=2)}{P(y_i=0)}\right) = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i.$$

Nende võrrandite sobival teisendamisel on võimalik leida tõenäosused vaatluse kategooriatesse sobimisele. Mudeli sobitamiseks kasutati selles uuringus paketi *nnet* funktsiooni *multinom*, mis tagastab ennustamisel automaatselt kõrgeima tõenäosusega kategooria. (Ripley ja Venables, 2025)

2.2.1 Tunnuste valik

Tunnused valiti mudelisse statistilise olulisuse alusel. Algne mudel hinnati kõikide saadaval olevate ennustajatega. Seejärel sooritati ükshaaval tõepärasuhte test algmudeli ja huvipakkuvate tunnusteta mudelite vahel. Suurima statistiliselt ebaolulise p-väärtusega testile vastav tunnus eemaldati seejärel mudelist. Valikuprotsessi korrati kuni kõikide tunnuste p-väärtus oli väiksem kui 0.05. Tuleb märkida, et sellise valikuga on võimalik, et ühe kategooria jaoks on tunnuse parameeter ebaoluline, sest tõepärasuhte test hindab tunnuse mõju koondmudelis.

2.2.2 Mudeli kvaliteedi hindamine

Eelneva metoodikaga hinnatud mudelit võrreldi baasmudeliga, mis leiti kasutades *MASS* paketi *stepAIC* funktsiooni. Algne mudel hinnatakse kõikide tunnustega, kuid seejärel eemaldatakse tunnus, mis langetab Akaike informatsioonikriteeriumi väärtust kõige rohkem. Selle valem on:

$$AIC = 2k - 2\ln(L),$$

kus k on mudeli parameetrite arv ja L on mudeli tõepära. Valikuprotsess lõppeb, kui ühegi tunnuse eemaldamisel enam AIC ei vähene. Selline valik tasakaalustab mudeli sobivuse ja parameetrite arvu ning ei nõua, et mudelisse kaasatud tunnused oleksid statistiliselt olulised. (Venables ja Ripley, 2002)

Mudeli ennustusvõime uurimiseks leiti mõlema mudeli parameetrid juhuslikult valitud vaatlustega treeninghulga põhjal, mis moodustas algandmestikust 80%. Ülejäänud 20% vaatlusi moodustasid testhulga, mida mudeli parameetrite hindamisel arvesse ei võetud. Hulkadesse jaotamisel kasutati kihistava tunnusena väljalangevust, et tagada gruppide osakaalude sarnasus kogu andmestikuga. Treeninghulgal leitud mudelit rakendati testhulga vaatlustel ning arvutati diskreetsed kategoorilised ennustused. Seejärel võrreldi mudeleid ennustuste ja tegelike väljalangevuse kategooriate põhjal arvutatud näitajatega.

Mudeleid võrreldi täpsuse ja ROC kõvera (*Receiver Operating Characteristics*) aluse pindala AUC põhjal (*Area Under Curve*). Täpsus on defineeritud kui korrektsete ennustuste arvu osakaal kõikide vaatluste arvust. ROC kõver binaarse tunnuse korral kirjeldab korrektsete positiivsete ennustuste määra funktsioonina valepositiivsete määrast. AUC on pindala selle kõvera all ning selle maksimaalne väärtus on 1, kirjeldamaks perfektse ennustusvõimega mudelit. Seevastu väärtus 0.5 on samaväärne võrdjuhusliku ennustamisega. Sellest madalam AUC tähendab seda, et mudeli ennustusvõime on halvem mündiviskega otsustamisest.

Kuna multinomiaalses mudelis on ennustataval tunnusel rohkem kui kaks faktori taset, laiendati ROC kõvera ja AUC definitsiooni. Seda on võimalik teha meetodiga, mida kirjeldavad Hand ja Till (2001). Uuritakse üht tunnuse taset korraga, võrreldes seda ülejäänud tunnuste vastu. Makro-AUC ehk keskmine AUC on seejärel defineeritud järgmise võrdusega:

$$M = \frac{2}{c(c-1)} \sum_{i<j} \hat{A}(i,j),$$

kus c on tasemete arv. Summeeritav väärtus on

$$\hat{A}(i,j) = \frac{\hat{A}(i \mid j) + \hat{A}(j \mid i)}{2},$$

kus $\hat{A}(i \mid j)$ on tõenäosus, et klassi j juhuslikult valitud vaatluse hinnanguline tõenäosus kuuluda klassi i on madalam juhuslikult valitud klassi i vaatluse tõenäosusest. $\hat{A}(j \mid i)$ on vastupidine tõenäosus. Hand ja Till (2001) näitasid oma artiklis, et see näitaja taandub binaarse tunnuse korral standardse AUC definitsioonile. Selle metoodika põhjal koostati ka *pROC* paketi funktsioon *multiclass.roc*, mida selles uuringus rakendatakse.

3 Analüüs

Selles töös kasutatud andmed koguti Tartu Ülikooli õppeinfosüsteemi andmebaasist. Kokku oli andmestikus infot 1207 tudengi kohta, kes olid immatrikuleeritud õppeaastatel algusega 2019-2023. Nende tudengite õppeteekondi käsitleti eraldiseisvate vaatlustena vastaval sisseastumisel arvutatud tunnustega. Ühe õpingute vaatluse moodustasid sisseastumisele või erialavahetusele järgnenud samal erialal õpitud semestrite andmed. 41 tudengi kohta oli kirjeldatud mitut erinevat õppeperioodi kas uuritavate õppekavade vahelise erialavahetuse või mitmekordse sisseastumise tõttu. Erinevate algustega õpingute vaatlusi oli kokku 1252.

Enne analüüsi eemaldati andmestikust 85 vaatlust. Neist 55 jäeti välja puuduva matemaatika eksami tulemuse, 17 eksmatrikuleerimise põhjuse ning 13 uuringu ulatusest väljaspool õpitud eriala tõttu. Eksmatrikuleerimise põhjustest jäeti välja need, mis ei olnud seotud õppetööga: oma soov majanduslikel põhjustel, õppeteenustasu tähtajaks tasumata jätmine ning üliõpilase surm. Analüüsis kasutati 1167 vaatlust 1139 tudengi õpingutest.

Selles peatükis kirjeldatakse uuritavate tunnuste statistilisi omadusi. Nendeks olid sugu, eriala, akadeemiline puhkus, matemaatika eksami tulemus, positiivsele tulemusele sooritatud EAP-de määr ja aritmeetiline keskmine hinne. Seejärel hinnatakse mudelid ja võrreldakse nende ennustusvõimet. Olulisuse nivooks selles uuringus valiti $\alpha = 0.05$.

3.1 Kirjeldav analüüs

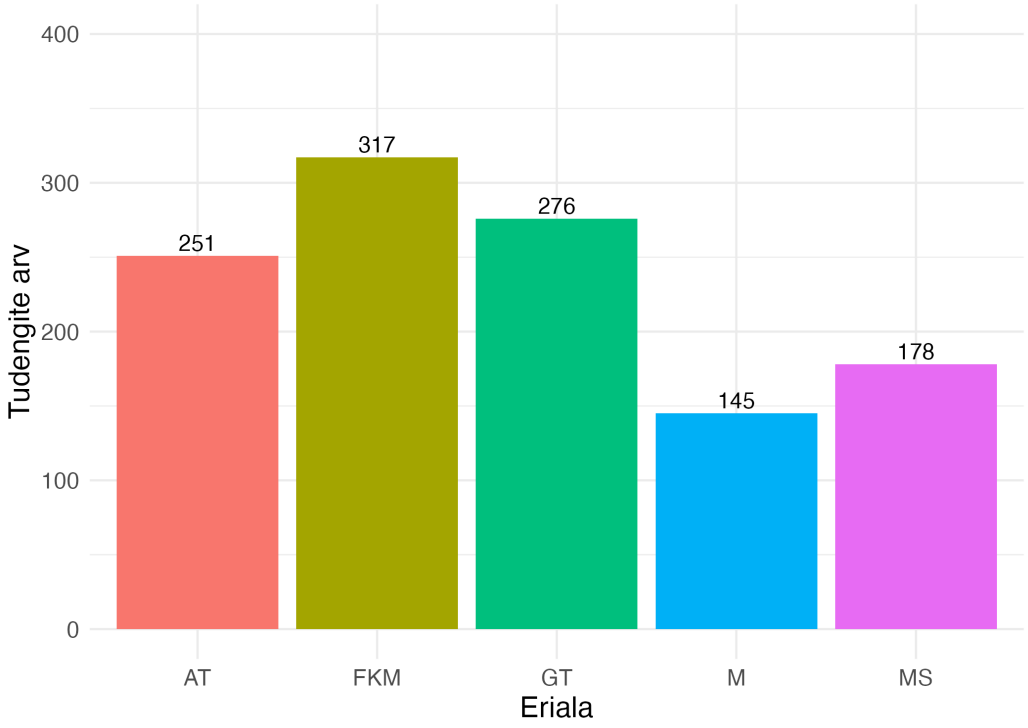
3.1.1 Sugu

Uuritavast 1139 tudengist 602 olid mehed ja 537 naised. Meestudengite osakaal oli 0.529. Selle osakaalu usaldusvahemik usaldusnivool $1 - \alpha = 0.95$ oli $(0.499, 0.558)$. Meestudengitest 11 alustasid õpinguid kahel ning 4 tudengit kolmel korral. Nais-

tudengite seas oli kaks sisseastumist 9 inimesel.

3.1.2 Eriala

Uuriti tudengite sisseastumist viiele LT valdkonna õppekavale. Arvutitehnika (AT) erialal alustasid õpinguid 251, füüsika, keemia ja materjaliteaduse (FKM) erialal 317, geenitehnoloogia (GT) erialal 276, matemaatika (M) erialal 145 ning matemaatilise statistika (MS) erialal 178 tudengit. Õppekavade jaotus on toodud joonisel 1.



Joonis 1: Tudengite arv viiel uuritaval erialal.

3.1.3 Akadeemiline puhkus

Uuritud õppeteekondadest 35 jooksul võttis tudeng akadeemilise puhkuse. Nende seast neljal korral võeti akadeemiline puhkus vähemalt kaks korda. Puhkuse võt-

mise põhjusteks olid kolmel korral kaitseväes teenimine, 19 korral oma soov ja 16 korral tervislikud põhjused. Akadeemilisele puhkusele suunduti 0.03 osal kõikidest õpingutest usaldusintervalliga $(0.021, 0.041)$.

Andmestikus polnud akadeemilist puhkust kirjas nendel tudengitel, kes võtsid selle ajateenistuses osalemiseks esimesel õppeaastal. Andmetöötluse käigus leiti aga 126 tudengit, kelle esimene õpitud semester toimus aasta peale immatrikuleerimist. See viitas kaitseväes teenimisele, aga täpsemate andmeteta polnud võimalik selles veenduda. Kuna esimesel õppeaastal ajateenistuse läbinud tudengite õppeteekond oli samaväärne õpingute alustamisega sisseastumisaastast üks aasta hiljem, siis ei arvestatud nendega mudeli jaoks tunnuse loomisel. Erandiks olid need tudengid, kes võtsid õpingute jooksul uuesti akadeemilise puhkuse.

3.1.4 Matemaatika eksami tulemus

Kitsa matemaatikaeksami sooritasid 68, laia eksami 1077 tudengit. Eksamitulemuste kirjeldavad statistikud on toodud tabelis 1. Laia eksami sooritajatest 36 tudengit saavutasid maksimumtulemuse.

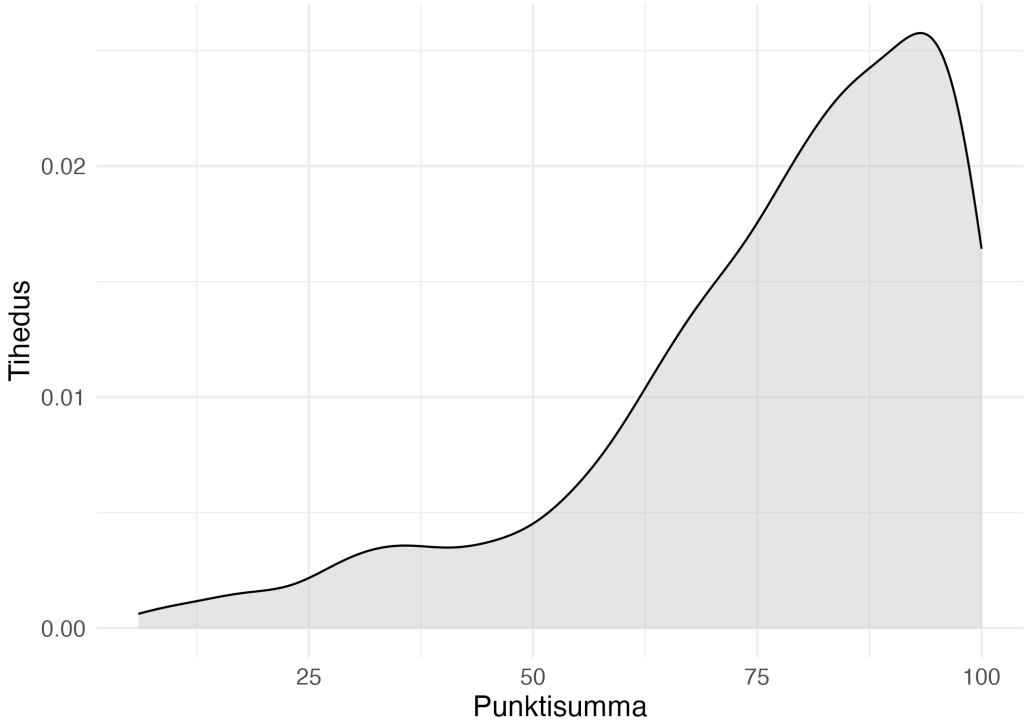
Tabel 1: Matemaatika eksami tulemuste kirjeldavad statistikud.

Eksam	Sooritajate arv	Keskmine	Standardhälve	Min	Max
Kitsas	68	67.37	24.27	16	99
Lai	1077	76.73	19.99	6	100

Kuuel tudengil oli andmestikus mõlema eksami punktisumma, mis kajastub ka tabelis, kuid sellele eripärale ei leitud andmetest selgitust. Ükski nendest tudengitest ei läbinud sisseastumist kahel korral. Mudeli hindamisel võeti mõlema tulemuse olemasolul kasutusele laia eksami punktid.

Võimalik põhjendus puuduvate matemaatika eksami tulemustele oli vaid 2 tudengil, kes kandideerisid olümpiaadiga. Ülejäänud 51 korral polnud andmestikus tunnuseid, mis selgitaks matemaatika eksami puudumist.

Matemaatika eksami tulemused võeti mudelis kasutusele kahe tunnusega: punktisumma ja eksami liik. Kokku oli eksamitulemuste keskmine 76.49 punkti standardhälbega 20.05. Usaldusvahemik oli $(75.34, 77.64)$. Tulemuste jaotuse tihedusgraafik on joonisel 2.



Joonis 2: Matemaatika eksami tulemuste tihedusgraafik.

3.1.5 Positiivsele tulemusele sooritatud ainepunktide määr

Ainepunktide eduka läbimise määr arvutati õppeteekonna esimese kahe õpitud semestri põhjal. Positiivsele tulemusele õpitud ainepunktide arv jagati registreeritud ainepunktide arvuga, et esindada tudengi toimetulekut võetud koormusega. Kui tudeng õppis vaid ühe semestri, siis arvutati tunnus olemasoleva semestri põhjal.

Positiivsele tulemusele sooritatud ainepunktide ja registreeritud ainepunktide kokkuvõtvad statistikud esimesel kursusel on tabelis 2. Esimene õppeaasta lõppes 107

korral nulli positiivsele tulemusele läbitud EAP-ga, millest 96 korral oli õpitud vaid üks semester.

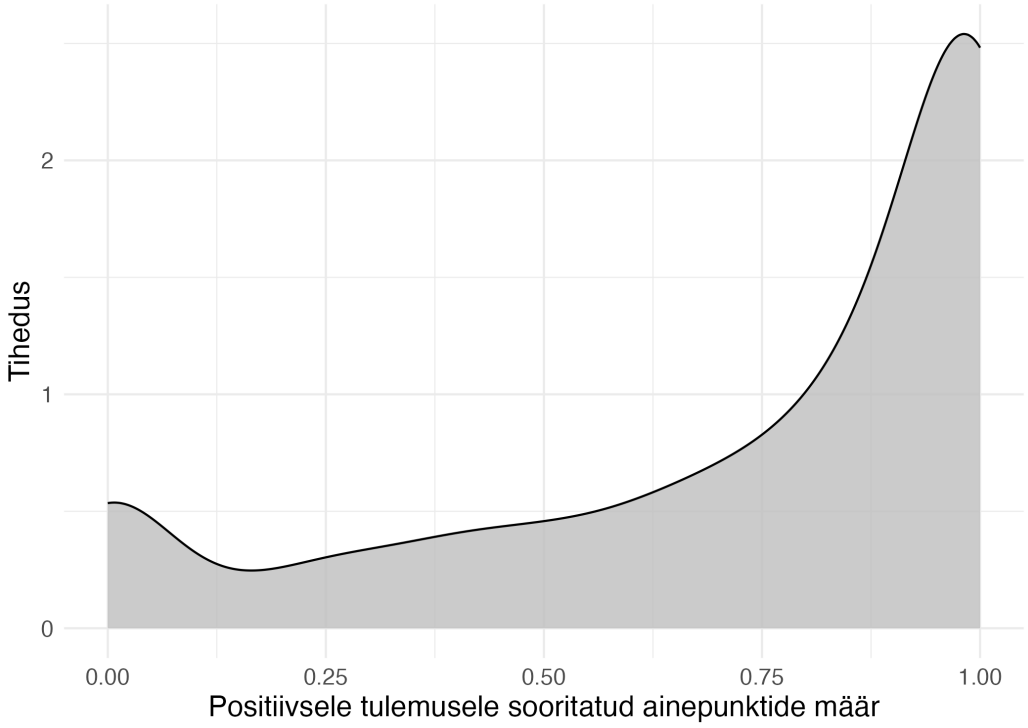
Tabel 2: Positiivsete ja registreeritud EAP-de kirjeldavad statistikud.

Tunnus	Keskmine	Mediaan	Standardhälve	Min	Max
Positiivsed EAP-d	36.77	45	18.64	0	99
Registreeritud EAP-d	47.28	49	14.46	3	99

Andmestikus oli ka 264 vaatlust, mille korral esimesel aastal registreeriti vähem kui 42 ainepunkti. Neist kahte selgitab tudengi suundumine akadeemilisele puhkusele ning 197 korral oli õpitud üks semester. 65 tudengit võtsid aga andmete põhjal kahe semestriga vähem ainepunkte kui nõutud, kellest 28 jätkasid õpinguid kolmandal semestril. Osakoormusega õppimises ei olnud andmete põhjal võimalik veenduda, kui tudengi eksmatrikuleerimise põhjus seda ei maininud. Osakoormusega õppimine ei väljendunud nendel vaatlustel ka hilisemas eksmatrikulatsioonis.

Kuna käesoleva tunnuse eesmärk oli kirjeldada õppeedukust esimesel kursusel vastavalt tudengi võetud koormusele, siis jäeti algsed vaatlused andmestikus muutmata. Samuti ei arvestatud akadeemilise puhkusega: registreeritud ainepunktide positiivselt sooritamine kajastab ka nende tudengite õppeedukust võetud koormusega esimesel õppeaastal.

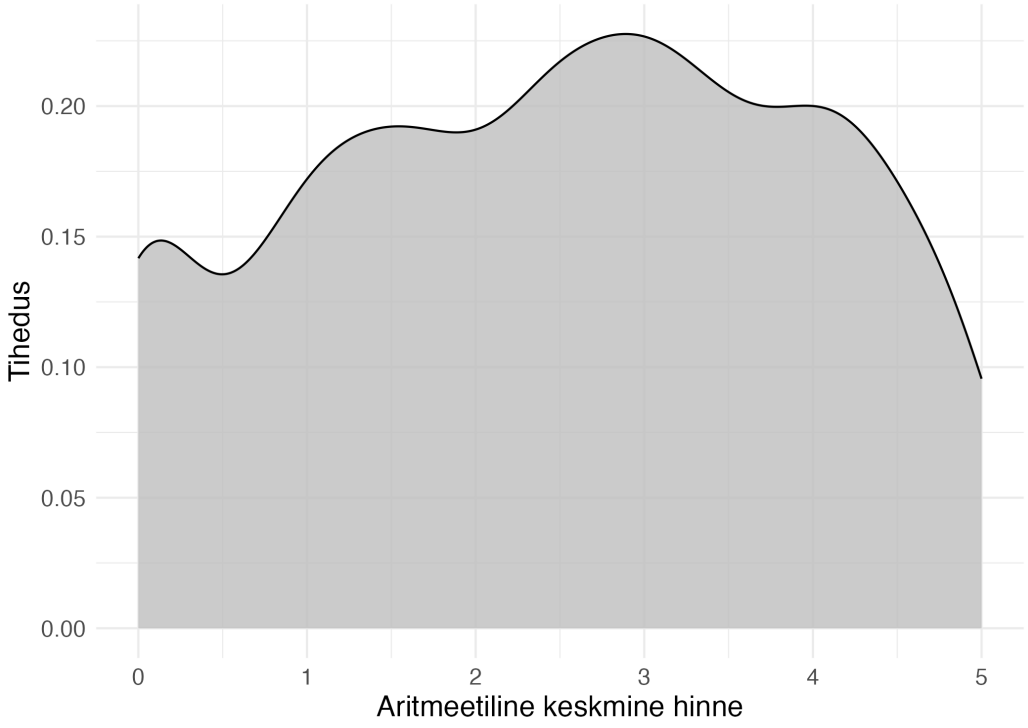
Positiivsele tulemusele sooritatud ainepunktide määra jaotus on joonisel 3. Selle tunnuse aritmeetiline keskmine oli 0.730, mediaan 0.884 ja standardhälve 0.330. Jaotus on kallutatud kõrgema positiivse EAP määra poole, kuid nulli lähedal on märgatav osa vaatlustest.



Joonis 3: Positiivsele tulemusele sooritatud ainepunktide määra tihedusgraafik.

3.1.6 Aritmeetiline keskmine hinne ülikoolis

Aritmeetiline keskmine hinne arvutati kõigi ühel erialal õpitud semestrite põhjal. Tudengite keskmine hinne õpingute jooksul oli 2.482 usaldusvahemikuga $(2.397, 2.566)$. Standardhälve oli 1.469. Tudengite keskmise hinde jaotus oli enamasti ühtlane, kuid alates hindest 4 kahanes kõrgema tulemusega tudengite arv märgatavalt (vt joonis 4).



Joonis 4: Aritmeetilise keskmise hinde tihedusgraafik.

3.1.7 Väljalangevus ja eriala vahetus

Viimase vaadeldud semestri seisuga õppisid 422 tudengit. Edukalt lõpetatud õppeteekondi oli 249, katkestatud 361 ja erialavahetusega lõppenud 135.

Nominaalajaga lõpetati õpingud 207 korral, neist üks tudeng lõpetas nelja semestriga ning kaks tudengit viie semestriga. Ühe aasta üle nominaalaja võtsid õpingute lõpetamiseks 41 tudengit ning ühel kulus õpingute edukaks lõpetamiseks 9 semestrit.

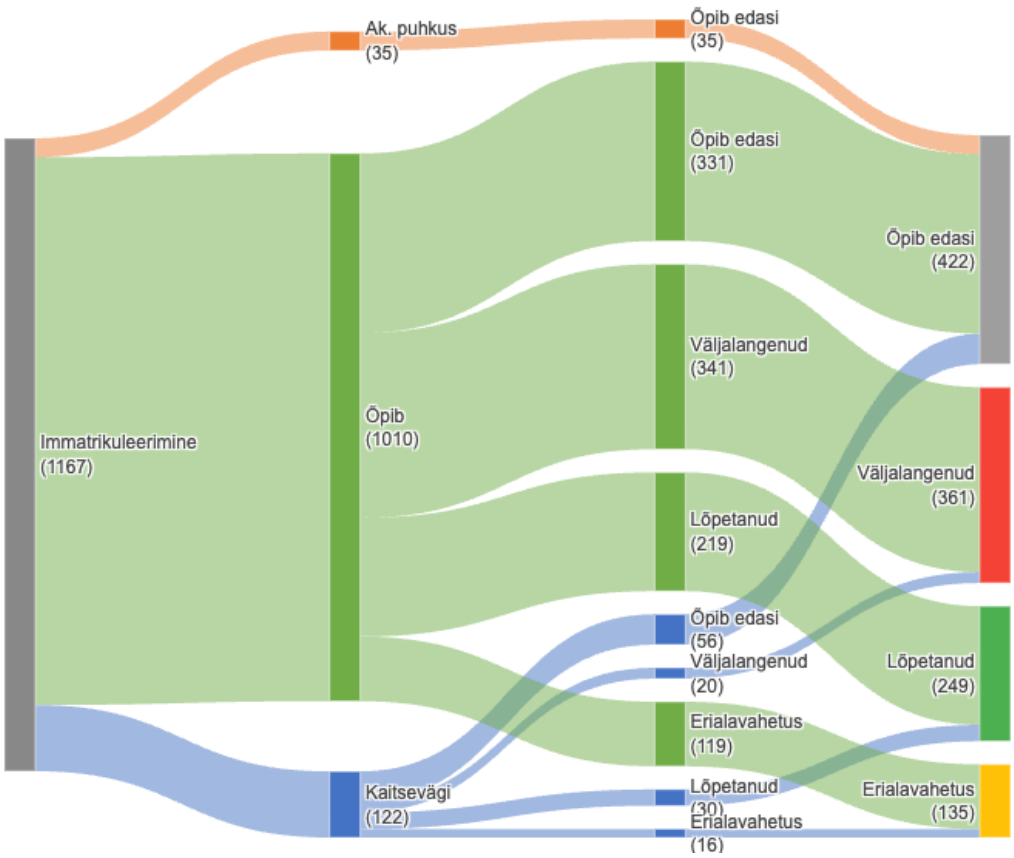
Vaatlusperioodi lõpus õppivatest inimestest 16 olid ületanud nominaalaja. Lisaks 33 inimest olid õppinud 6 semestrit, aga polnud eksmatrikuleeritud viimase semestri seisuga. 14 tudengit õppisid neljandat aastat ning 2 õppisid üheksandal semestril. Õpingute kestuste sagedus õppivate tudengite seas on tabelis 3. Need õpingud

võivad lõppeda hilisemal ajal katkestamise või eriala vahetusega. Selles uuringus määratakse need tudengid ühte baaskategooriasse koos edukate lõpetajatega. See valik võib aga mõjutada analüüsi tulemusi.

Tabel 3: Õppivate tudengite õpingute kestus semestrites.

Semestrite arv	Tudengeid
1	5
2	226
3	7
4	131
5	4
6	33
7	5
8	9
9	2

Tudengite õppeteekonna koos akadeemilise puhkuse ning kaitseväes teenimisega võtab kokku joonis 5. Ajateenistuses oli neli inimest, kes märgiti vaid akadeemilise puhkuse võtjana sellel joonisel. Kõik akadeemilise puhkuse võtjad olid veel õppimas vaatlusperioodi lõppseisuga.



Joonis 5: Tudengite õpingute teekond.

3.2 Mudelite analüüs

Peale puuduvate andmete ja ebasobilike vaatluste eemaldamist jäi mudelite hindamiseks andmestikku 1167 õppeteekonda. Nendest 80% jaotati treeninghulka ja 20% testhulka. Kihistamiseks kasutati uuritavat väljalangevuse tunnust, et kategooriate osakaal hulkades vastaks tervele andmestikule.

Kirjeldava analüüsi käigus leiti, et kõik akadeemilise puhkuse võtjad jätkasid õpinguid. See tähendab, et puhkuse tunnus ennustas perfektselt baaskategooriat 0, mis tekitas probleeme mudelite hindamisel. Tunnuse kaasamisel mudeli ennustusvõime kasvas tehislikult, sest parameetri hinnang oli absoluutväärtuselt suur. Selle

piirangu tõttu otsustati akadeemiline puhkus mudelite koostamisel välja jätta, et ka nende vaatluste korral oleks võimalik teiste tunnuste ennustusvõimet ja mõju usaldusväärselt uurida.

3.2.1 Multinomiaalse regressiooni mudelid

Multinomiaalsest mudelist eemaldati järgmised tunnused:

1. eksamiliik ($p = 0.7952$)
2. sugu ($p = 0.6635$)
3. eriala ($p = 0.6863$)

Mudeli koefitsientid ja Waldi testi tulemused on tabelis 4. Erialavahetuse alamudelis ei olnud statistiliselt eristatav aritmeetilise keskmise hinde ning väljalangevust ennustavas mudelis matemaatika eksami tulemuse parameetri hinnang. Kindlalt saab eristada EAP määra mõju: mida suurema osa esimesel kursusel võetud koormusest tudeng positiivsele tulemusele läbib, seda tõenäolisem on, et ta ei vaheta eriala ega katkesta õpinguid.

Tabel 4: Multinomiaalse mudeli koefitsiendid.

Tulemus	Tunnus	Koefitsient	SE	z	p
vahetus	(vabaliige)	−0.1560	0.6111	−0.255	0.7986
	matemaatika eksam	0.0204	0.0070	2.938	0.0033
	EAP määr	−3.0657	0.7814	−3.924	<0.001
	keskmine hinne	−0.1990	0.1615	−1.232	0.2179
katkestamine	(vabaliige)	3.4540	0.4412	7.828	<0.001
	matemaatika eksam	0.0008	0.0050	0.170	0.8651
	EAP määr	−3.9162	0.6996	−5.598	<0.001
	keskmine hinne	−0.5935	0.1605	−3.699	<0.001

Testhulga vaatluste põhjal arvutatud täpsus oli 0.7362 ja Makro-AUC oli 0.737. See mudel ei ennustanud aga ühtegi eriala vahetajat: 16 ennustati baaskategooriasse

ning 11 õpingute katkestajaks. Kontrollmudeli hindamisel kaasati samad tunnused, seega mitteoluliste tunnuste lisamine ei oleks parandanud ka AIC põhjal mudeli sobivust.

3.2.2 Binomiaalse logistilise regressiooni mudelid

Kuna multinomiaalne mudel ei olnud võimeline ennustama eriala vahetamist, uuriti seda ja väljalangevust eraldi binaarse logistilise regressiooniga. Selleks eraldati algsest andmestikust kaks uut andmestikku. Esimeses olid õpingud, mille lõppseis oli õppimine, lõpetamine või eriala vahetamine. Selliseid vaatlusi oli kokku 806. Teise andmestiku moodustamisel eemaldati algandmestikust eriala vahetanud tudengite vaatlused. Kasutati 1032 õppeteekonnaga andmestikku. Mõlemad valimid jaotati treening- ja testhulkadeks vastavalt multinomiaalse mudeli metoodikale.

Erialavahetust ennustava binaarse mudeli koostamisel eemaldati järgmised tunnused:

1. aritmeetiline keskmine hinne ($p = 0.9214$),
2. eksamiliik ($p = 0.6069$),
3. sugu ($p = 0.4711$),
4. eriala ($p = 0.3219$).

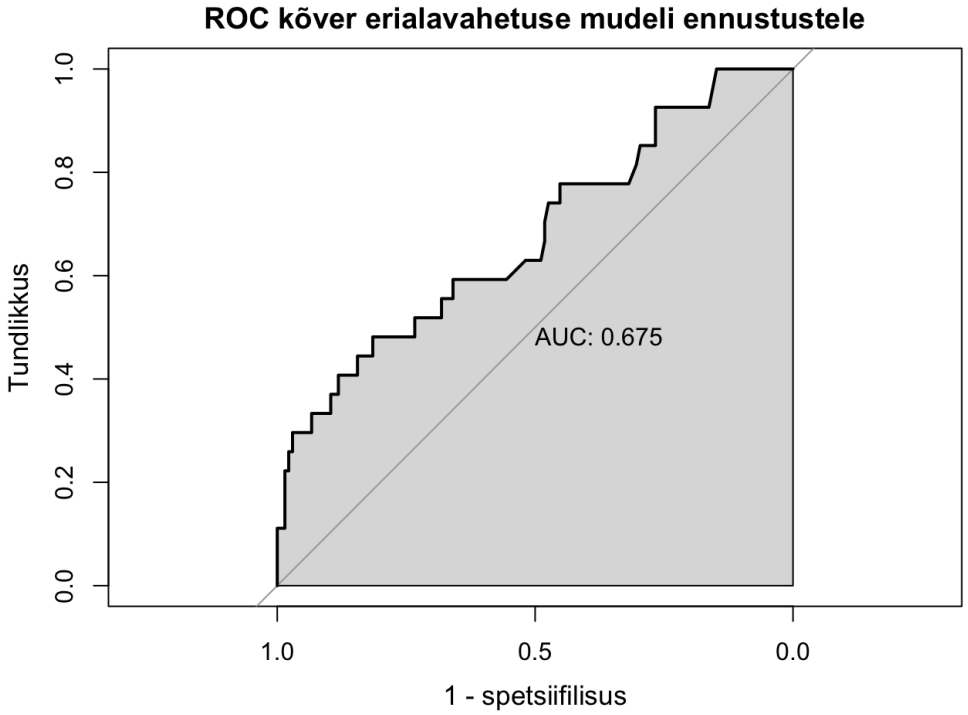
Mudelisse kaasati matemaatika eksami tulemus ($p = 0.0438$) ja positiivsete ainepunktide määr ($p < 0.001$). Parameetrid ja Waldi testi tulemused on tabelis 5. Ainus statistiliselt oluline parameetri hinnang oli ainepunktide määr: suurem edukalt läbitud osa võetud koormusest langetab tudengi tõenäosust vahetada eriala.

Kontrollmudelisse kaasati AIC põhjal samad tunnused, mitteolulised tunnused seega mudeli sobivust ei parandanud. Selle mudeli täpsus oli testhulga vaatluste põhjal 0.858 ja AUC oli 0.6754. Põhjalikumal uurimisel leiti, et mudel ennustas testhulga 27 erialavahetajast vaid 6 õigesti. 135 lõpetanud või õppivast vaatlusest vaid

Tabel 5: Erialavahetuse binomiaalse mudeli koefitsiendid.

Tunnus	Koefitsient	SE	z	p
(vabaliige)	0.7190	0.6088	1.181	0.2376
matemaatika eksam	0.0130	0.0067	1.936	0.0528
EAP määr	−4.0975	0.5103	−8.029	<0.001

2 prognoositi valesti, mis põhjustas ka kõrge täpsuse. Mudeli prognooside põhjal leitud ROC kõver on joonisel 6.



Joonis 6: ROC kõver ja AUC erialavahetuse mudeli ennustustele

Väljalangevuse mudeli koostamisel eemaldati tunnused järgmises järjekorras:

1. sugu ($p = 0.9387$),
2. matemaatika eksami tulemus ($p = 0.8320$),
3. eksamiliik ($p = 0.5127$),

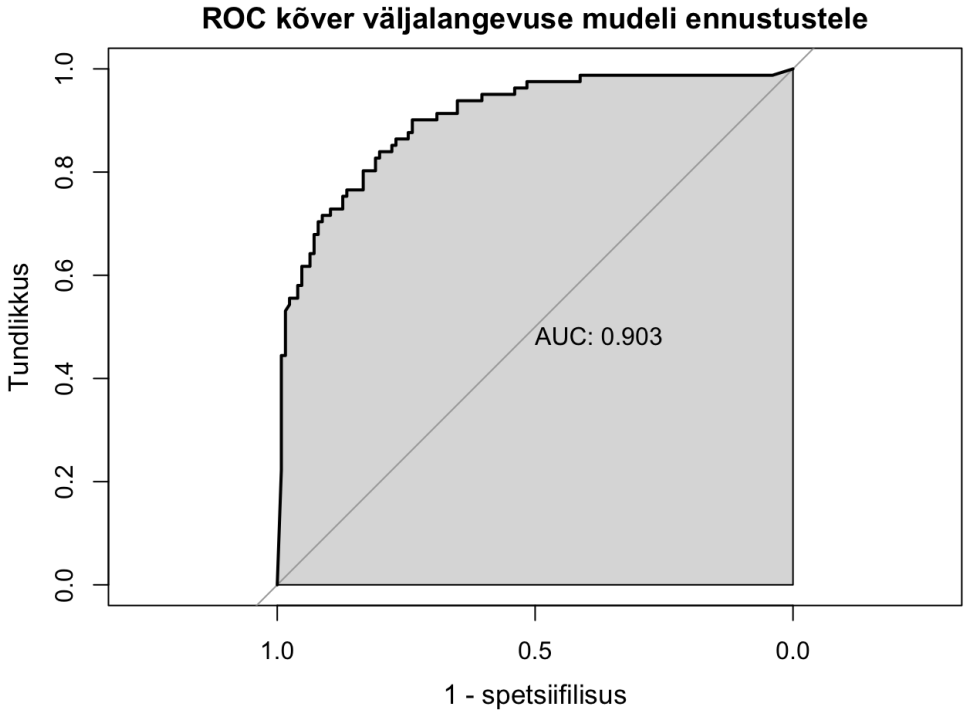
4. eriala ($p = 0.2528$).

Mudelisse kaasati positiivsete ainepunktide määr esimesel kursusel ja aritmeetiline keskmine hinne (mõlemal juhul $p < 0.001$). Parameetrite hinnangud ja Waldi testi tulemused on tabelis 6. Kõrgema positiivsele tulemusele läbitud ainepunktide määraga tudengitel on oluliselt madalam tõenäosus katkestada õpingud. Sama kehtib aritmeetilise keskmise hinde kohta.

Tabel 6: Väljalangevuse binomiaalse mudeli koefitsiendid.

Tunnus	Koefitsient	SE	z	p
(vabaliige)	3.4670	0.3372	10.283	<0.001
EAP määr	−3.9191	0.7206	−5.439	<0.001
keskmine hinne	−0.5665	0.1585	−3.575	<0.001

Kontrollmudelisse kaasati samad tunnused, mudeli sobivus ei paranenud mitteoluliste tunnuste kaasamisel. Testhulga ennustustel arvutatud täpsus oli 0.8261 ja AUC oli 0.9025. Mudel ennustas õigesti 59 õpingute katkestajat 81 seast. Ühtlasi määrati 14 baaskategooria 126 vaatlusest katkestajaks. Ennustuste põhjal leitud ROC kõver on joonisel 7.



Joonis 7: ROC kõver ja AUC väljalangevuse mudeli ennustustele

3.3 Arutelu

Multinomiaalse mudeli hindamisel selgus, et kolme uuritavat õppeteekonna lõppseisundi kategooriat ei ole võimalik ühe mudeliga kirjeldada kasutades saadavalolevaid tunnuseid. Mudel ei olnud võimeline ennustama ühtegi eriala vahetajat. Selle tõttu uuriti neid kategooriaid eraldi kahe binaarse mudeliga.

Väljalangevuse mudelist tuvastati oluliste mõjuteguritena esimesel õppeaastal positiivsele tulemusele läbitud ainepunktide määr ja aritmeetiline keskmine hinne (mõlemal $p < 0.001$). Mudel saavutas testhulgas AUC 0.903. Need leiud on kooskõlas varasemate uuringutega: Chen (2013) leidis samuti, et ainepunktide läbimine ja keskmine hinne on väljalangevuse olulised ennustajad. Läbitud ainete arvu mõju esimese kursuse väljalangevusele oli oluline ka Casanova *et al.* (2023) uuringus.

Erialavahetuse binaarse mudeli ennustusvõime jäi nõrgaks. Mudel saavutas AUC 0.675 ning ennustas testhulga 27-st eriala vahetajast õigesti vaid 6. Nende lõppseisudega vaatlused võisid olla eristamatud saadaval olevate tunnuste põhjal. Chen (2013) leidis, et eriala vahetajatel on sarnaselt lõpetajatega kõrgem keskmine hinne, mille tõttu selle tunnuse ebaolulisus mudelis oli oodatav. Ennustusvõimet võivad parandada Pedersen ja Nielsen (2024) metoodikale sarnased küsimustikupõhised andmed, millega saab hinnata tudengi motivatsiooni ja huvisid. Leitud mudel kaasas ainepunktide määra ja matemaatika eksami tulemuse, kuid eksami parameetrit ei saanud Waldi testi põhjal veenvalt hinnata ($p = 0.053$). Oluline mõjutegur erialavahetusele oli selle analüüsi põhjal esimese kursuse positiivsete ainepunktide määr.

Nii multinomiaalse kui ka eraldi binaarsete mudelite korral valiti tõepärasuhte testiga ja Akaike informatsioonikriteeriumiga samad tunnused lõppmudelisse. AIC-ga oli võimalik kaasata mudelisse tõepärasuhte testi järgi ebaolulisi tunnuseid, kui need parandasid mudeli sobivust. Nende välja jätmine tähendab, et pole alust eeldada, et statistiliselt mitteolulised tunnused oleksid parandanud mudeli ennustusvõimet.

Kirjeldava analüüsi tulemusena selgus, et kõik 35 akadeemilisele puhkusele suundunud tudengit olid õppimas viimase vaadeldud semestri seisuga. See tulemus oli vastuolus Naylor, Cox ja Cakitaki (2023) uuringuga, kus vähem kui kolmandik puhkuse võtjatest jätkas õpinguid või lõpetas eriala. Akadeemilise puhkuse võtmine võib selle tulemuse põhjal olla olulise mõjuga tudengi õpingute jätkamisele. Õpingute lõppseis ei olnud aga nendel vaatlustel teada ning mõni õppeteekond võis hiljem lõppeda katkestuse või eriala vahetusega. Akadeemilise puhkuse mõju tuleb uurida pikema ajaperioodi põhjal, et kaasata ka lõplike staatustega õpinguid.

Analüüsis oli kaks peamist piirangut. Esiteks lisati viimase vaadeldud semestri seisuga õppivad tudengid edukate lõpetajatega samasse kategooriasse, kuigi osa neist võis hiljem õpingud katkestada või eriala vahetada. See kallutas mudeleid baas-

kategooria ennustamise suunas. Lisaks võisid mudeli parameetrite hinnangud olla konservatiivsemad. Teiseks ei olnud võimalik hinnata õppeinfosüsteemi andmete põhjal tudengite motivatsiooni ja huvisid. Nende andmetega oleks edasises analüüsis võimalik lähemalt uurida, mis põhjustab madalat õppeedukust. Lisaks võib nende andmete kasutamine olla vajalik eriala vahetamise prognoosimisel.

Kokkuvõte

Käesoleva töö eesmärk oli tuvastada olulised mõjutegurid väljalangevusele ja erialavahetusele Tartu Ülikooli loodus- ja täppisteaduste valdkonnas. Lisaks hinnati, kas statistiliselt mitteoluliste tunnuste lisamine parandab ennustusvõimet kummagi kategooria korral. Uuriti 1167 õppeteekonnal põhinevat valimit viielt LT õppekavalt aastatel 2019-2023 immatrikuleeritud tudengitest.

Väljalangevuse ja erialavahetuse ennustamine ühise multinomiaalse regressiooni mudeliga ei õnnestunud. Mudel ei olnud võimeline ennustama erialavahetajaid, mille tõttu uuriti õpingute lõppseise eraldi binaarsete mudelitega.

Binaarsest logistilise regressiooni mudelist tuvastati väljalangevuse oluliste mõjuteguritena esimesel õppeaastal positiivsele tulemusele läbitud ainepunktide määr ja kõikide õpitud semestrite aritmeetiline keskmine hinne. Tunnuste mõju oli oodatav: mida suurem osa esimesel kursusel võetud koormusest läbiti edukalt ja mida kõrgem oli keskmine hinne, seda madalam oli õpingute katkestamise tõenäosus. Leitud mudeli ennustusvõime oli rahuldav, saavutades testhulgal AUC väärtuse 0.903.

Erialavahetust ei olnud võimalik selle uuringuga veenvalt ennustada. Ainsaks statistiliselt oluliseks mõjuteguriks osutus positiivsete ainepunktide määr ning mudeli ennustusvõime oli nõrk ($AUC = 0.675$). Erialavahetuse uurimiseks võib vaja minna motivatsiooni ja teisi õppeinfosüsteemi väliseid tegureid kirjeldavaid andmeid.

Uurimusel oli kaks peamist piirangut. Vaatluste seas olid viimase vaadeldud semestri seisuga õppivad tudengid, kelle õpingud võisid hilisemal vaatlusel lõppeda erialavahetuse või välja langemisega, kuid lisati baaskategooriasse koos edukate lõpetajatega. Lisaks puudusid andmed tudengite motivatsiooni ja enesehinnangu kohta. Edasistes uuringutes on soovitatav uurida pikemat vaatlusperioodi ning täiendada õppeinfosüsteemi andmeid küsimustikupõhiste andmetega.

Hoolimata nendest piirangutest on leitud tulemused praktilise väärtusega. Esimese õppeaasta positiivsele tulemusele läbitud ainepunktide määr ning aritmeetiline

keskmine hinne on jälgitavad tunnused, mille põhjal saab sekkuda, kui väljalangevus nende põhjal tõenäoliseks muutub.

Kasutatud allikad

Aulck, L., R. Aras, L. Li, C. L'Heureux, P. Lu ja J. West (2017). *Stemming the Tide: Predicting STEM attrition using student transcript data.* University of Washington Information School. DOI: 10.48550/arXiv.1708.09344.

Belser, C. T., M. A. Shillingford, A. P. Daire, D. J. Prescod ja M. A. Dagley (2018). "Factors Influencing Undergraduate Student Retention in STEM Majors: Career Development, Math Ability, and Demographics". *Professional Counselor* 8 (3), lk. 262–276. DOI: 10.15241/ctb.8.3.262.

Casanova, J. R., A. Castro-López, A. B. Bernardo ja L. S. Almeida (2023). "The Dropout of First-Year STEM Students: Is It Worth Looking beyond Academic Achievement?" *SUSTAINABILITY* 15 (2), lk. 1253–1264. DOI: 10.3390/su15021253.

Chen, X (2013). *STEM Attrition: College Students' Paths into and out of STEM Fields. NCES 2014-001.* National Center for Education Statistics, Insitute of Education Sciences ja U.S. Department of Education. URL: https://files.eric.ed.gov/fulltext/ED544470.pdf.

Eurostat (2023). *Early school leavers down to 10% in 2022.* URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20230523-2 (vaadatud 04.04.2025).

– (2024). *Early leavers from education and training.* Eurostat. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Early_leavers_from_education_and_training#Highlights (vaadatud 04.04.2025).

Hand, D. J. ja R. J. Till (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems". *Machine Learning* 45, lk. 171–186. DOI: 10.1023/A:1010920819831.

Haridus- ja Teadusministeerium (2022). *Haridus- ja Teadusministeeriumi valdkondade analüüs. (2021. aasta tulemusaruannete lisa)*. Haridus- ja Teadusministeerium. URL: https://www.hm.ee/sites/default/files/documents/2022-10/tulemusaruande_lisa_-_valdkondade_analuus_0.pdf (vaadatud 04.04.2025).

Kreegipuu, T ja I. Jaggo (2017). *LTT erialadel õppimine Eesti kõrghariduses*. Haridus- ja Teadusministeerium. URL: https://www.hm.ee/sites/default/files/documents/2022-10/ltt_erialad.pdf.

Naylor, R., S. Cox ja B. Cakitaki (2023). "Personalised Outreach to Students on Leave of Absence to Reduce Attrition Risk". *The Australian Educational Researcher* 50 (2), lk. 433–451. DOI: 10.1007/s13384-021-00503-2.

OECD (2022). *Education at a Glance 2022: OECD Indicators*. DOI: 10.1787/3197152b-en.

Pedersen, J. V. ja M. W. Nielsen (2024). "Gender, self-efficacy and attrition from STEM programmes: evidence from Danish survey and registry data". *Studies in Higher Education* 49 (1), lk. 47–61. DOI: 10.1080/03075079.2023.2220702.

Prosper, E. (2024). "Retention and Attrition in U.S. STEM Education with the Help of Computer Technology and Curriculum Development". *International Journal of Scientific Research and Management* 12 (12), lk. 3847–3865. DOI: 10.18535/ijsrm/v12i12.el07.

Ripley, B. ja W. Venables (2025). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. DOI: 10.32614/CRAN.package.nnet.

Varga, B., K. Fodor ja R. Szilágyi (2023). "Analysis of the Factors Affecting Student Dropout at the University of Miskolc". *Multidiszciplináris tudományok* 13 (2), lk. 261–274. DOI: 10.35925/j.multi.2023.2.23.

Venables, W. N. ja B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Yu, R., H. Lee ja R. F. Kizilcec (2021). *Should College Dropout Prediction Models Include Protected Attributes*. DOI: 10.48550/arXiv.2103.15237.

Lisad. Analüüsi olulisem R-i kood

Nendes lisades on esitatud analüüsi olulisemad R-is kirjutatud koodilõigud.

Lisa 1. Andmestiku moodustamine ja vaatluste välistamine

Andmestik `andmed2` võtab kokku vaatlused (tudengi eraldiseisev õppeteekond). Lisaks eemaldatakse andmestikust vaatlused, mille korral puudus matemaatika eksami tulemus, eksmatrikuleerimise põhjus polnud seotud õppetööga või õpitud eriala jäi uuringu ulatusest välja. Algandmestik `andmed1` on pikal kujul andmestik, kus iga tudengi eraldiseisva õppeteekonna semestrid on eraldi ridadena välja toodud.

```
andmed2_koik <- andmed1 %>%
  group_by(ISIK2) %>%
  summarise(
    õpitud_semestreid = length(sem2),
    sugu       = SUGU[õpitud_semestreid],
    eriala     = OPPEKAVA_NIMETUS_PARANDATUD[õpitud_semestreid],
    puhkus     = as.numeric(any(puhkus == 1)),
    mat_eksam  = mat[õpitud_semestreid],
    eksamiliik = eksam[õpitud_semestreid],
    s1_EAP     = EAP_TOTAL[sem2==1] + ifelse(2 %in% sem2, EAP_TOTAL[sem2
↪ ==2], 0),
    s1_EAP_POS = EAP_POS[sem2==1]   + ifelse(2 %in% sem2, EAP_POS[sem2
↪ ==2],   0),
    ARITM_AVG  = sum(ARITM_AVG_AINED) / sum(AINEID_KOKKU),
    põhjus     = EKSMAT_POHJUS[õpitud_semestreid],
    .groups    = "drop"
  )

andmed2_koik <- andmed2_koik %>%
```

```
  mutate(EAP_määr = s1_EAP_POS / s1_EAP)

andmed2_koik <- andmed2_koik %>%
  left_join(väljalangevus, by = "ISIK2") %>%
  mutate(väljalangevus = ifelse(väljalangeja == 1, 2,
                          ifelse(erialavahetus == 1, 1, 0)))

andmed2_koik <- andmed2_koik %>%
  mutate(
    errors = as.integer(is.na(mat_eksam)) +
             as.integer(põhjus %in% c(
               "omal soovil majanduslike põhjuste tõttu",
               "õppeteenustasu tähtajaks tasumata jätmise tõttu",
               "üliõpilase surma tõttu"
             )) +
             as.integer(!(eriala %in% c("AT", "FKM", "GT", "M", "MS")))
  )

andmed2 <- andmed2_koik %>%
  filter(errors == 0) %>%
  drop_na(väljalangevus, sugu, mat_eksam, eksamiliik,
          EAP_määr, ARITM_AVG, eriala) %>%
  filter(is.finite(EAP_määr), is.finite(ARITM_AVG))
```

Lisa 2. Tunnuste valik tõepärasuhte testi alusel

Funktsioon `model_selection` sooritab tagurpidi sammuviisilise tunnuste valiku tõepärasuhte testi p-väärtuste alusel. Igal sammul leitakse algmudelist ja huvipakkuva tunnuseta mudelist tõepärasuhte testi p-väärtus. Eemaldatakse tunnus, mille p-väärtus on suurim ja ületab olulisuse nivood 0.05. Protsess kordub kuni kõikide allesjäänud tunnuste p-väärtus on alla olulisuse nivoo.

```
model_selection <- function(full_model) {
  logs  <- data.frame()
  model <- full_model
  repeat {
    tunnused <- attr(terms(model), "term.labels")
    if (length(tunnused) <= 1) break
    p_values <- sapply(tunnused, function(var) {
      uus <- update(model, as.formula(paste(". ~ . -", var)), trace =
↪ FALSE)
      lrt <- lrtest(model, uus)
      lrt[2, "Pr(>Chisq)"]
    })
    worst_p   <- max(p_values, na.rm = TRUE)
    worst_var <- names(which.max(p_values))
    if (worst_p > 0.05) {
      model <- update(model, as.formula(paste(". ~ . -", worst_var)),
                      trace = FALSE)
      logs  <- rbind(logs, data.frame(var = worst_var, p = worst_p))
    } else {
      break
    }
  }
  list(model, logs, p_values)
}
```

Lisa 3. Mudelite hindamine ja paralleelne AIC võrdlus

Mudeli hindamiseks jaotati andmestik treening- ja testhulkadeks suhtega 4:1, kasutades kihistava tunnusena väljalangevust. Algmudel hinnati kõikide kandidaattunnustega ning seejärel rakendati nii `model_selection` (tõepärasuhte test) kui ka `stepAIC` (AIC järgi valik), et võrrelda valitud tunnustehulki.

```
set.seed(123)
split    <- initial_split(andmed2, prop = 4/5, strata = väljalangevus)
training <- training(split)
test     <- testing(split)

full_multinom <- multinom(
  väljalangevus ~ sugu + mat_eksam + eksamiliik +
                  EAP_määr + ARITM_AVG + eriala,
  data = training, trace = FALSE
)

# Tõepärasuhte testi alusel
res_multinom    <- model_selection(full_multinom)
model_multinom  <- res_multinom[[1]]

# AIC alusel
step_multinom   <- stepAIC(full_multinom, direction = "backward",
                           trace = FALSE)

# Ennustused testhulgal
pred_class <- predict(model_multinom, test)
prob       <- predict(model_multinom, test, type = "prob")
acc        <- mean(pred_class == test$väljalangevus)
mauc       <- auc(multiclass.roc(test$väljalangevus, prob))
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Ken-Erik Aus,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Loodus- ja täppisteaduste valdkonna üliõpilaste väljalangevus", mille juhendaja on Mare Vähi, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Ken-Erik Aus
11.05.2026