TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Natalja Maksimova 192017IABM

IKT eriala üliõpilaste varajase väljalangemise ennustamine masinõppe meetodite abil

Magistritöö

Juhendaja: Olga Dunajeva

Phd

Kohtla-Järve 2021

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Natalja Maksimova

10.12.2021

Annotatsioon

Arvutiteaduse programmidest üliõpilaste kõrge väljalangemine on oluline ülemaailmne probleem. TalTech Virumaa kolledžis on juba aastaid seisnud silmitsi suure väljalangemise probleemiga esimese kursuse informaatikaüliõpilaste seas. Käesolevas töös kirjeldatakse uuringu tulemusi TalTech Virumaa kolledži Telemaatika ja arukate süsteemide eriala esmakursuslaste väljalangemise ennustamise mudelite konstrueerimisest struktureeritud ja struktureerimata andmete põhjal. Tähelepanu on kontsentreeritud üliõpilaste varajase väljalangemise analüüsimisele ja ennustamisele, põhinedes andmetele, mis on enne esimese semestri õppetulemuste avaldamist kättesaadavad. Ennustamiseks kasutati sisseastumiseprotsessi käigus kogutud andmed nagu (1) sisseastujate andmed koos sisseastumisevestluse tulemustega; (2) online küsitluse sisseastujate vabatahtlikud vastused ja (3) IKT temaatikaga seotud esseede tekstid. Ennustamise mudeleid ehitati masinõppe AdaBoost, logistilise regressiooni, Naive Bayes, osavähimruutude diskriminantanalüüsi (PLS-DA), otsustuspuu ja tugivektor-masina meetoditel.

Uuringu töökeskkonnaks on RStudio ja programmeerimiskeel R. Uuringu käigus oli selgitatud väljalangevust mõjutavad tegurid välja, kasutades neid tegureid hinnati ennustavaid mudeleid ja selle tulemusena töötati välja soovitused vastuvõtuprotsessi parandamiseks ja esmakursuslaste väljalangevuse vähendamiseks.

Töö parim mudel on tugivektor-masina mudel (ristvalideerimise meetodi keskmine F1 skoor on 0.7825), mis oli loodud leitud mõjukatel tunnustel: Keskhariduse lõputunnistuse keskmine hinne, Linna suurus („väike“), sisseastuja elukoht (Ida-Virumaa), vanus, essee leksikaalne rikkus ja sisseastumisvestluse punktid, kusjuures viimasel on väga nõrk mõju.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 51 leheküljel, 8 peatükki, 36 joonist, 12 tabelit.

Abstract

ICT Students Early Drop-Out Prediction Using Machine Learning

The high number of university dropouts from Computer Science programs is a big problem worldwide. The dropout rates of Informatics program freshmen have been persistently high for years in TalTech Virumaa College. This paper describes the results of the conducted research gotten from constructed TalTech Virumaa College's Telematics and Smart Systems program freshmen dropout rates prediction models based on structured and unstructured data. The paper's attention is mainly focused on analyzing and predicting college students' premature dropping out based on the data available before the publication of the first semester results. The following collected during the admission process data were used for prediction: (1) data received during the high school graduates' college interviews; (2) answers received from high school graduates in the voluntary online questionnaire and (3) essay texts related to ICT subject. Prediction models were built using machine learning methods: AdaBoost, decision trees, Naïve Bayes, Support Vector Machines, Partial Least Squares Discriminant Analysis (PLS-DA) and Logistic Regression.

The work environments used for research are RStudio and R programming language. The determinants that affect the dropout rate were found during the research and assessed using the prediction models. As a result, the paper introduces advice for improving the admission process as well as for decreasing the first-year dropout numbers.

The work's best-working model is a support-vector machine (cross-validation's F1 score average is F1=0.7825), which was created based on the following found influential features: high school diploma grade point average, town size („small"), high school graduate's place of residence (Ida-Virumaa), age, essay's lexical diversity and interview points, where the latter has very weak influence.

The thesis is written in Estonian and includes 51 pages, 8 chapters, 36 figures and 12 tables.

Lühendite ja mõistete sõnastik

Accuracy	*Accuracy*, täpsusmäär, mis näitab, milline oli õigete ennustuste osakaal kõikidest ennustustest
AdaBoost	*Adaptive Boosting*, masinõppe algoritm
CV	*Cross Validation*, ristvalideerimine
EAP	õppemahu arvestusühik, mis kasutatakse töömahu mõõtmiseks
flesch_kin	*Readability Score*, Flesch-Kincaidi teksti loetavuse hinnang
lexdiv	*Lexical Diversity*, teksti leksikaalse mitmekesisuse näitaja
LogR	*Logistic Regressioon*, logistiline regressioon, masinõppe algoritm
LOOCV	*Leave-One-Out Cross Validation*, jäta-üks-vahele ristvalideerimine
NB	*Naive Bayes*, Naiivne Bayes, masinõppe algoritm
PLS-DA	*Partial Least Squares Discriminant Analysis*, osavähimruutude diskriminantanalüüs (PLS-DA)
Rpart	*Recursive Partitioning*, masinõppe algoritm
SAIS	Eesti tudengitele haridusasutustesse sisseastumise infosüsteem
SKKH	semestri kaalatud keskmine hinne
SVM	*Support Vector Machines*, tugivektor-masinad, masinõppe algoritm
TalTech või TTÜ	Tallinna Tehnikaülikool
VIF	*Variance Inflation Factor*, dispersiooni mõju faktor
ÕIS	Õppeinfosüsteem

Sisukord

Jooniste loetelu

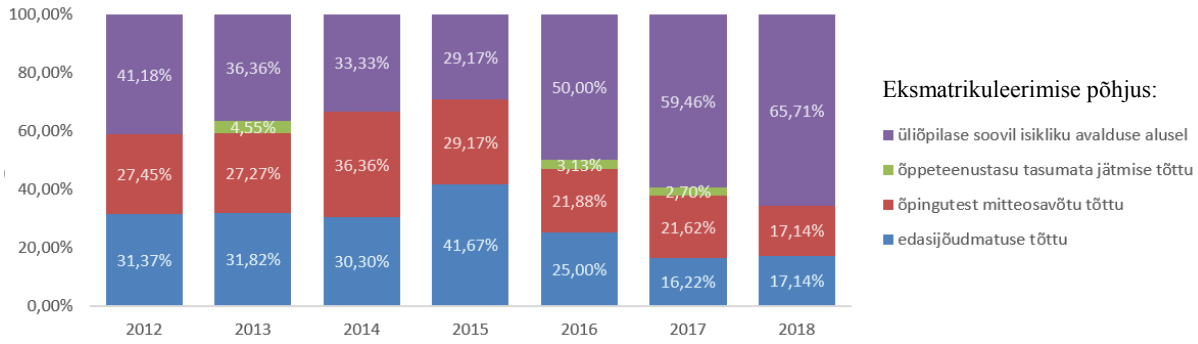
Tabelite loetelu

1. Sissejuhatus

Tänapäevases maailmas, kus valdavad digitaalsed lahendused, protsesside automatiseerimine uute masinõppel põhinevate tehnoloogiate kaudu, protsesside kaugjuhtimine, on IT-valdkonna erialad muutunud väga nõutuks. Tehnoloogia ja teadmiste väärtus kasvab majanduses iga aastaga. Prognoositakse, et 2024. aastaks kasvab nõudlus ainuüksi programmeerijate järele 61% ja moodustab 6400 töökohta [1, lk13].

Selle taustal on IKT-valdkonna üliõpilaste väljalangemine iga IKT-erialasid pakkuva ülikooli jaoks võtmeküsimus. Tulevase tööturu vajaduste rahuldamiseks on iga IKT-eriala pakkuv ülikool huvitatud sellistest üliõpilastest, kes esimesel õppeaastal ei lahku.

Tallinna Tehnikaülikooli Virumaa kolledži jaoks on see küsimus eriti terav seetõttu, et riigi- ja koolieksamite sooritamine ei ole sisseastumise eeldus. Vastuvõtu tingimus IKT-erialale oli kuni 2019. aastani ainult lõputunnistuse keskmine hinne, mis muutus 2012.-2018. aastatel alates 3.0 kuni 3.6. Nende erialade esmakursuslaste väljalangevus on ülejäänud kolledži erialadega võrreldes alati kõrge. Aastatel 2012 - 2018 oli esmakursuslaste keskmine väljalangevus 43%. Joonisel 1 on kujutatud kolledžist varajase väljalangemise põhjuste jaotus igal aastal. Kõige sagedasem eksmatrikuleerimise põhjus „oma soovil" oli määratud keskmiselt 46%, edasijõudmatuse tõttu 26.5% ja õpingutest mitteosavõtu tõttu 25% juhtudel. Kolledži õppeosakonna läbi viidud täiendavast küsitlusest selgus, et õpinguid katkestanud üliõpilased ei kujutanud ette ega tundnud hästi valitud eriala ning nende õppimismotivatsioon langes mõni kuu pärast õpingute algust. Neid üliõpilasi ei ole otstarbekas sellele erialale võtta, et üliõpilane ei kaotaks aastat ja kolledž ei kaotaks üliõpilasi.



Joonis 1. Esmakursuslaste eksmatrikuleerimise põhjused.

Õppeedukus on üks olulisemaid väljalangemist mõjutavaid tegureid. On juba tõestatud, et teades ainult kaalutud keskmist hinnet ja sooritatud EAP-de arvu, on 90% täpsusega võimalik ennustada üliõpilase kolledžist lahkumise võimalust [2]. Selles töös vaatleb autor andmeid, mida on võimalik saada kolledžisse sisseastumise ajal, et keskenduda varajase, enne õppimise algust väljalangemise ennustamisele. Riskirühma kuuluvate üliõpilaste varajane tuvastamine sisseastumisel võimaldab pakkuda neile lisatoetust juba õpingute alguses.

Alates 2019. aastast on kolledž võtnud tarvitusele meetmeid, et vähendada IT-erialade esmakursuslaste suurt väljalangevust. Vastuvõtutingimuseks sai kohustuslik vestlus, IT-erialade õppejõudude eeltutvus sisseastujaga, mille käigus selgitatakse välja mõlema poole nõudeid ja nägemused. Lõputunnistuse keskmise hinde piirmäär tõusis 3.6-ni ja 2021. aastal 3.7-ni. Lisaks rakendatakse mentorlusprogrammi, esmakursuslaste toetamine kolledži töötajate poolt.

2020. aastal koostati kolledžis IKT-erialale sisseastujale prooviküsimustik. E-küsimustik sisaldas küsimusi motivatsiooni, telemaatika ja arukate süsteemide eriala nägemuse ja kogemuste kohta.

Peale selle kirjutasid juba telemaatika ja arukate süsteemide erialale sisse astunud üliõpilased esseesid teemal „Minu nägemus tulevasest erialasest tööst”, mis aitasid välja selgitada nende motivatsioonid ja ettekujutused.

Selles töös loetakse kirjutatud esseesid motivatsioonikirjaks, mida saab kandideerijatele pakkuda kirjutamiseks sisseastumisel. Seega saadakse teavet tulevase üliõpilase kohta erinevatest allikatest: üliõpilase täidetud SAISi andmed, kolledži vestluse tulemused, kolledži elektroonilise küsimustiku vastused ja kirjutatud essee.

Selle töö eesmärk oli kogu eelpool kogutud info läbitöötamine, eri tüüpi info efektiivsuse analüüsimine väljalangevuse vaatevinklist ning Virumaa kolledži IKT-eriala esmakursuslaste väljalangemise mudelite koostamine, kasutades andmeid, mida on võimalik saada enne kolledžisse astumist.

Eesmärk:

Luua meetod TalTech Virumaa kolledži telemaatika ja arukate süsteemide eriala esmakursuslaste väljalangemise ennustamiseks struktureeritud ja struktureerimata andmete põhjal.

Ülesanded:

1. Analüüsida Eesti kõrgkoolide IKT-eriala tudengite varajase väljalangemise ennustamise teemal avaldatud uurimis- ja magistritöid.
2. Läbi viia sisseastumise etapil ja enne esimese semestri lõppu saadavate telemaatika ja arukate süsteemide eriala esmakursuslaste struktureeritud ja struktureerimata andmete statistiline analüüs ning leida tudengite väljalangemist mõjutavad faktorid.
3. Koostada võimalikult efektiivne mudel üliõpilaste varajase väljalangemise ennustamiseks, kasutades erinevaid klassifitseerimise algoritme.
4. Analüüsida tulemusi ja pakkuda soovitused sisseastumisprotseduuri efektiivsuse tõstmiseks väljalangemise vähendamise jaoks.

Autor pakkus välja mitu hüpoteesi, mida selles töös tõestatakse või lükatakse ümber.

1. Sisseastumisvestlus on üliõpilaste väljalangevuse seisukohalt tõhus meede.
2. Küsimustik võimaldab välja selgitada üliõpilaste teadlikkust valitud suunast ja see aitab kaasa prognoosimismudelile.
3. Motivatsioonikirja lisamine sisseastumisel suurendab üliõpilaste väljalangemise ennustamise tõhusust.
4. Otsuse tegemiseks piisab eelnevalt kogutud täielikust teabest (üldandmed, vestlus, ankeet ja motivatsioonikiri).

Vestluse, küsimustiku või essee tõhusust väljalangemise ennustamisel hinnatakse vastavate tunnuste ja uuritava tunnuse vahelise seose olemasolu ja olulisuse järgi. Nende tunnuste üksik- ja kogupanus prognoosimudelitesse aitab hinnata nende tõhusust väljalangevuse prognoosimisel. Hüpotees meetodi tõhususe kohta võetakse vastu, kui

vastavate tunnuste alusel loodud mudeli varajase väljalangevuse ennustamisvõime ületab 70%.

See uuring on korraldatud järgmiselt. II osas vaadatakse väljalangevuse teemal tööde ülevaade. III osas kirjeldatakse uuringu andmed, IV jaotises – kasutatav metoodika. V osas antakse uurimuslik analüüs, mõjutavate tunnuste leidmine ja tunnuste selekteerimine, VI osas – masinõpe rakendamise tulemused. VII jaotises esitatakse töö järeldused ja VIII on kokkuvõte.

2. Probleemi taust

2.1 Esmakursuslaste väljalangevuse teemal Eestis kirjutatud tööde eelvaade

Esmakursuslaste väljalangemise teema on aktuaalne mitte ainult Virumaa kolledži jaoks. Sarnaseid uuringuid on läbi viinud ja viivad läbi mitmed eri riikide ülikoolid, et mõista oma õppeasutuses üliõpilaste väljalangemise põhjuseid, ning jagavad saadud tulemusi erinevatel konverentsidel. Nende uuringute eesmärk on leida meetodid väljalangemisriskiga üliõpilaste varajaseks leidmiseks ja neile õigeaegse õppeasutusepoolse abi osutamiseks. Selles vallas on uuringuid teinud ka Eesti suurimad ülikoolid. Järgmistes lõikudes on toodud nende uurimiste tulemuste analüüs.

2015. aastal Tartu Ülikooli psühholoogia instituudi uurimisrühma tehtud uuringu „Haridustee valikud ning õpingute katkestamise asjaolud Eesti kõrghariduses" tulemusena tehti subjektiivsete asjaolude analüüsi [3], millega üliõpilased põhjendasid õppekava valikut ja õpingute peatumist ja katkestamist. Uuritavate rühma kuulusid üliõpilased erinevatest Eesti ülikoolidest. Uuringu tulemuste kirjeldavas artiklis on öeldud, et enam kui pooled üliõpilased, kes esmakordselt alustavad kõrgharidusõpinguid, mitte ei lõpeta, vaid katkestavad õpingud enne õppekava täitmist. Mõned tudengid katkestavad oma õpinguid, et asendada need teise tegevusega. Õppimise katkemine on kogu Eesti kõrghariduse probleem, mis puudutab kõiki õppetasandeid ja õppesuundi. Uurimus näitas, et õppetööga seotud emotsionaalsed probleemid ja kõrgkooli integratsiooni tase [4] mõjuvad õppetöö katkestamisele. Siinjuures selgus, et vastuvõtutestide roll akadeemilise edu prognoosimisel on piisav vaid kõrgkoolis õppimise esimesel aastal.

Samal aastal on ilmunud Tallinna Tehnikaülikooli ja Tartu ülikoolide uurimisrühma artikkel „First-year dropout in ICT studies" [5] infotehnoloogia valdkonna esmakursuslaste väljalangevuse uuringust. Uuringu tulemusena sai selgeks väljalangevuse mõjutegurite loetelu. Oli kinnitatud, et väljalangevuse küsimuses on olulised järgmised tegurid: tudengi isiklik motivatsioon, teenitud ainepunktid, varasem

õppimine, õpingute ootused. Järgmiste tegurite, nagu vanus, sugu, õpingute ajal töötamine, sõprade arv IKT-valdkonnas väljalangevusele mõju ei olnud kinnitanud. Muuhulgas autorid leidsid, et õpinguid katkestanud õpilastel olid madalamad punktid matemaatika riigieksamil. Artiklis on näidatud IKT-valdkonna esmakursuslaste väljalangevuse keskmine protsent 32.2%.

Õigeaegne teave riskirühmast üliõpilaste kohta võimaldab õppejõududel/ülikoolil suunata oma tähelepanu nende toetamisele. Mõned autorid soovitavad teha pärast esimest semestrit küsitlust [5], teised - sisestada e-õppe süsteemi algoritm, mis oskaks ennustada tudengi õppeedukust tema tulemuste ja kursuse külastatavuse järgi ja mille abil on võimalik ennetada tudengite väljalangemist. Moodle'i platvormi andmeid kasutas Oskar Liblik oma 2016. a. kaitstud magistritöös „Klassifitseerimise algoritmi abil e-õppe süsteemis tudengite väljalangemise ennetamine informaatika aine näitel" [6]. Hetkel puudub e-õppe süsteemides tudengite õppevõimekuse ning saavutuste põhjal järelduse tegemise võimalus ühe semestri registreeritud kursuste lõikes. Õppeplatvormil Moodle on võimalus ainult ühe kursuse tulemuste analüüsiks ja tegevuste jälgimiseks, see on moodul Completion Progress, mis visualiseerib tööde sooritamise ja nende hindamise progressi (Joonis 2).

Joonis 2. Moodul Completion Progress Moodle'is.

Nagu märkab selle töö autor, Moodle'i logid (tegevusaruanded) annavad palju mitteaktuaalset ja üleliigset infot, mis raskendab nende läbitöötlust ja edasi masinõppe algoritmi kasutamist. Andmete hoidmine serveris selle muutmatul kujul ei ole võimalik, sest võtab palju ressursse.

Struktureeritud ja korraldatud tudengite andmed on olemas igas ülikoolis oma õppeinfosüsteemis (edasi ÕIS). Selles süsteemis on olemas andmed sisseastumise ja õppimise etappidel. Brenda Uga oma 2017. aastal kaitstud bakalaureusetöös „TTÜ tudengite väljalangemise ennustamine: tõenäosuse arvutamine masinõppemeetodite abil ning tulemuste kuvamine veebirakenduses" [7] tegi oma uuringuid juba ÕIS-i andmete põhjal. TTÜ IKT tudengite väljalangemise ennustamise algoritmidena olid valitud

klassifitseerimise algoritmid, nagu otsustuspuu, otsustusmets, logistiline regressioon, mitmekihiline pertseptron jms. Mudelite ennustustäpsused olid saadud vahemikus 60-95%. Kõige olulisemaks väljalangevust mõjutavaks faktoriks oli selles töös nimetatud tudengil õppimise ajal kogutud EAP-de arvu. Uuringus kasutatud andmete põhjal on kõige tähtsamad tunnused ennustamiseks: EAP-de arv, keskmine hinne, eelmine haridustase ja sisseastumisaasta. Sugu, kodakondsus ja sünniaasta ei oma tähtsust väljalangemise ohu tekkimisel. Tudengite väljalangemise riskigruppi kuulumise tõenäosuse hindamiseks võib kasutada masinõppe algoritmide abil loodud mudeleid.

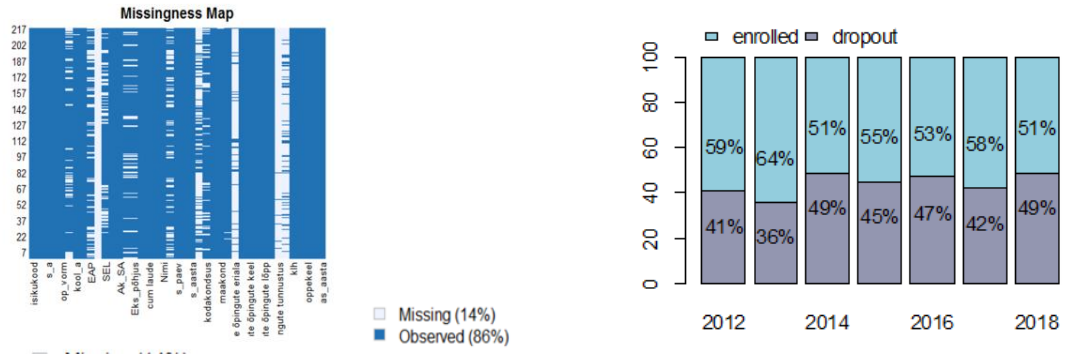
Iga ülikooli huvitab küsimus, kas tulevase tudengi akadeemilist edukust võib vastuvõtuprotsessi käigus ennustada. Tartu Ülikooli Toria Ketevan „Predicting Academic Performance From Admission Scores and Application Data –A Case Study" 2019. a. magistritöös oli hinnatud sisseastumispunktide efektiivsust ning oli koostatud Software Engineering magistriprogrammi tudengite akadeemiliste tulemuste prognoosimismudel, analüüsides sisseastumisdokumente andmekaevandamise metoodikate abil [8]. Autor leidis, et sisseastumishinde arvutamine, mis põhineb ainult bakalaureuseõppe keskmisel hindel ei ole piisav ja edaspidi omab nõrka positiivset mõju keskmisele akadeemilisele tulemusele. Töö tulemusena selgus, et efektiivsem (ennustustäpsusega 43%) on vastuvõtusüsteem, kus vaadatakse motivatsioonikirja, motivatsioonikirja plagiaati, õppekeele (inglise keel) oskust, vanust ja riiki. Selles süsteemis soovitab töö autor pöörata rohkem tähelepanu motivatsioonikirja hindamisele ja vähem bakalaureuseõppe keskmisele hindele.

Väljalangevuse uurimistööde analüüs näitas, et autoritel puudub ühine lähenemisviis mudelile, mis oleks kõigile ülikoolidele efektiivne nii tudengite väljalangevuse kui ka nende akadeemiliste tulemuste ennustamisel. Ülikoolide tudengite andmed tulevad erinevatest allikatest, on tihti üleliigsed ja nõuavad tõsist eeltöötlust. Andmete eeltöötluse protsessi lihtsustamiseks oleks hea luua intelligentsed võimalused andmete kogumise automatiseerimiseks ja standardsel vormil esitamiseks. Masinõppe meetodeid käsitletavates töödes on õppeedukus kõige olulisem faktor ja ennustusmudelid annavad häid tulemusi keskmiselt üle 80%. Vastuvõtuprotsessi käigus kogutud andmete põhjal väljalangevuse ennustamismudelid nõuavad arendamist.

2.2 TalTech Virumaa kolledži tehtud uuringu ülevaade

Virumaa kolledži töörühm tegi ka esmakursuslaste väljalangevuse uuringu ÕIS-i andmetel [2]. Uuring viidi läbi 2012.-2018. aastate kolledži IT-eriala üliõpilaste andmestikul, mis koosnes 367 objektist. Autorid jagasid ÕIS-i andmed kaheks osaks: (1) tudengi andmed enne õppimise algust ja (2) I semestri õppimise andmed.

Esimene osa (sisseastuja andmed enne õppimise algust, edasi andmestik A) sisaldab üldteavet sisseastuja ja tema hariduse kohta, mida vastuvõtukomisjon sisestas SAIS-i. Andmestik A oli koostatud ÕIS-is olemasolevate ja vastuvõtukomisjoni kogutud andmetest. Komisjoni kogutud andmed jõudsid autorini ühtset vormingut jälgimata ja nõudsid palju töötlemist: ühtsesse vormingusse viimist, puuduvate väärtuste asendamist, tunnuse osadeks jagamist ja uute tunnuste loomist. Joonis 3 illustreerib puuduvate andmete arvu ja IKT-eriala esmakursuslaste väljalangemise protsenti.



(a) (b)

Joonis 3. 2012. – 2018. aastate kolledži IKT eriala üliõpilaste andmed enne õppimise algust: (a) puuduvate andmete osakaal, (b) igaaastane väljalangemise protsent (värvitud tumedama värviga).

Mudelite ehitamiseks oli kasutatad Weka programmi klassifitseerimise algoritmid: otsustuspuud, logistiline regressioon, Naiivne Bayes, tugivektor-masin ja närvivõrgu algoritm. Oli leitud, et kasutades ainult tudengi andmeid enne õppimise algust väljalangemise klassifitseerimisprobleemis, maksimaalse ennustustäpsusega (Accuracy 70%) mudel on ehitatud *Matemaatika riigieksami tulemus* ja *Keskhariduse lõputunnistuse keskmine hinne* andmetel. Oli kinnitatud ka eelneva haridustaseme (kutsehariduse) mõju.

Teises osas semestri õppimise andmetel, I semestri EAP[1] arv ja SKKH[2], konstrueeritud mudel annab ennustustäpsust 90%. Ei olnud kinnitatud, et ainete hinnete lisamine tõstab prognoositäpsust.

1 EAP - õppemahu arvestusühik, mis kasutatakse töömahu mõõtmiseks

2 SKKH – semestri kaalatud keskmine hinne

3. Andmed

Andmestik A oli kirjeldatud punktis 2.2. Käesolevas töös on kasutatud erinevatest allikatest ja erineva struktuuriga 2019. ja 2020. aastate andmeid. Andmestik B sisaldab 2019. ja 2020. aastate sisseastujate ja vestluse andmed. Andmestikus C on 2019. ja 2020. aastate esseede andmed ja andmestikus D – 2020. aasta küsitluse andmed (vt Tabel 1). Käesoleva uuringu sihtmuutuja *y*, väljalangemise indikaator, on binaarne tunnus väärtusega 1 (tudeng on eksmatrikuleeritud esimese õppeaasta jooksul) või 0 (tudeng ei ole eksmatrikuleeritud esimese õppeaasta jooksul).

Tabel 1. Töös kasutatud andmekogumite karakteristikud.

Andmete karakteristikud		**Andmestik B**	**Andmestik C**	**Andmestik D**
suurus	objektid n	89	67	30
	atribuudid X (originaalfailides)	15	1	54
sihttunnus	sihtmuutuja y	1 – on eksmatrikuleeritud esimese õppeaasta jooksul (edasi positiivne klass); 0 – ei ole eksmatrikuleeritud (edasi negatiivne klass)		
	tasakaalustatud	(45%/55%) (1/0)		
	tasakaalustamata		(33%/ 67%) (1/0)	(37%/ 63%) (1/0)
sõltumatud tunnused	teisendused/uued tunnused	+	+	+
	puuduvad väärtused			+
	kokku atribuute pärast eeltöötlust	16	5	49
	pidevad	2	5	1
	diskreetsed	3	-	1
	nominaalsed	2	-	12
	järjestatud	2	-	28
	binaarsed	7	-	7

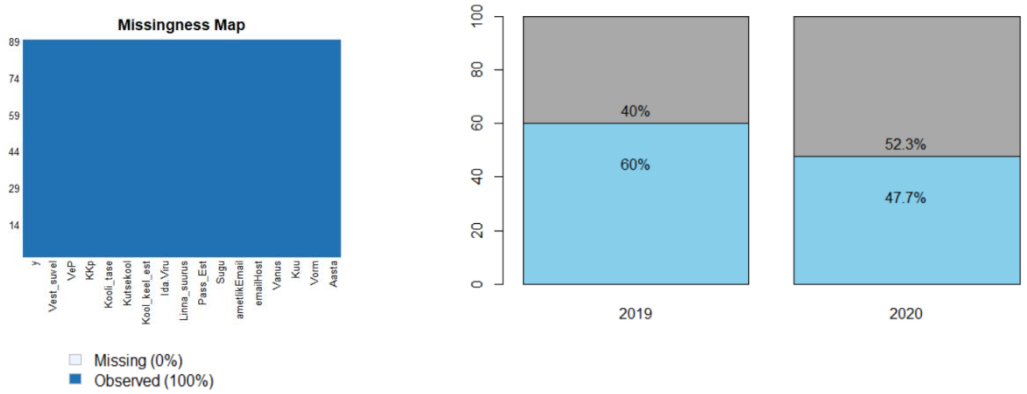
	Kas oli leitud multikollineaarsus	+	–	+

Esimene õppeaasta tähendab akadeemilise kalendri järgi sügissemestri õpingukava esitamise alguse tähtajast kuni üliõpilase lõppenud õppeaasta soorituste alusel järgmiseks õppeaastaks koormuse arvutamise tähtajani. Andmete markeerimise ja eeltöötluse protsess on mahukas. Markeerimine on vaja sihttunnuse *y* moodustamise jaoks ja eeltöötluse protsess – analüüsi ja modelleerimise jaoks. Punktides 3.1 – 3.3 on kirjeldatud lähemalt andmed ja eeltöötluse protsess. Andmete eeltöötlus oli teostatud R abil.

3.1 2019. ja 2020. aastate sisseastujate andmed koos vestlusega (andmestik B)

Alates 2019. aastast kogutakse tunnistuse hindeid ja üldteavet sisseastujate kohta SAIS-i süsteemi ilma kõrgkooli osaluseta. Kolledž hakkas SAIS-ilt andmeid saama päringuga kolledžisse vastuvõtutingimuste alusel. Andmed sisaldavad sisseastujate üldandmeid, nagu sugu, kodakondsus; kontaktandmed: telefon, aadress, e-post; keskhariduse andmed: õppeasutuse nimetus, lõputunnistuse keskmine hinne, õppekeel ja andmed vestluse kohta: vestluse keskmised tulemused[3] ja vestluse kuupäevad. Andmetes puudub teave matemaatika riigieksami tulemuste kohta.

Andmed tulid päringule vastavalt ühtsel kujul ja ilma puuduvate väärtusteta (Joonis 4a).



(a) (b)

Joonis 4. 2019. ja 2020. aastate kolledži IKT eriala üliõpilaste andmed enne õppimise algust: (a) puuduvate andmete osakaal, (b) igaaastane väljalangemise protsent (värvitud tumehalli värviga).

[3] Virumaa kolledži vestluse käik ja teemad on esitatud Lisas 3.

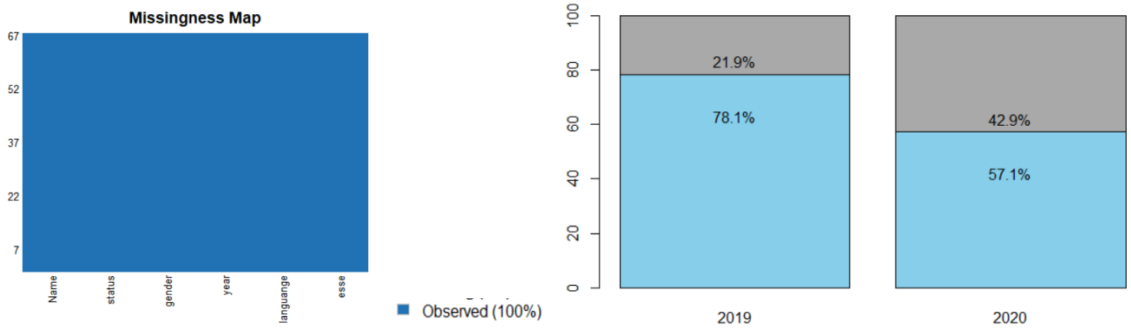
2020. aastal oli üks kõrgemaid väljalangevusi (Joonis 4b), mida mõjutas varajane (oktoobri lõpust alguse saanud) ja pikk üleminek distantsõppele. See selgus isiklike vestluste ja ankeedi analüüsi tulemusena [9].

Kuna matemaatika riigieksami tulemus oli esimeses uuringus oluline tunnus, siis autor otsustas luua matemaatika riigieksami tulemusega seotud nominaalset tunnust „Kooli tase matemaatika riigieksami tulemuste pingereas" kooli nimetuse ja kooli 2019[4]. ja 2020[5]. aasta edetabelite järgi. Pingerea koostamiseks võeti arvesse eksamitulemused: eesti keele (emakeelena või teise keelena), kitsa matemaatika, laia matemaatika ja inglise keele riigieksamid.

Andmestik B on andmestikuga A sarnane, need kogutakse andmetest kolledžisse sisseastumise ajal. Tunnuste, tehtud teisenduste ja uute tunnuste üksikasjalik kirjeldus asub Lisas 2.

3.2 2019. ja 2020. aastate esseed (andmestik C)

Andmed on 2019. ja 2020. aastate üliõpilaste õppeaines „Sissejuhatus erialasse" kirjutatud esseed teemal „Minu nägemus tulevasest erialasest tööst". Essee kirjutati I semestri lõpus, kirjutajate hulgas ei ole üliõpilasi, kes katkestas oma õppimist enne seda või ei kirjutanud esseed mingil teisel põhjusel. Kirjutatud esseesid on kokku 67, mis moodustab 75,3% sisseastunutest 2019. ja 2020. aastatel. Joonisel 5b on 2019. ja 2020. aasta väljalangemiste protsentide võrdlus.



(a) (b)

Joonis 5. 2019. ja 2020. aastate kolledži IKT eriala tudengite esseed: (a) puuduvate andmete osakaal, (b) igaaastane väljalangemise protsent (värvitud tumehalli värviga).

[4] https://infogram.com/koolide-edetabel-2019-mj-1h7g6k1q8p704oy
[5] http://cf.datawrapper.de/bhPiJ/5/

Esseede kogum oli autoril võetud Moodle`i õpikeskkonnast ja kontrollitud plagiaadituvastussüsteemi Urkund abil, mis on integreeritud Moodle`isse. Esseede vormistus ei olnud rangelt reglementeeritud ja algul autor ei saanud täielikult programmeerida teksti võtmise protsessi, sest esseede vormistamine erines väga palju. Edaspidiseks kasutamiseks oli protsess lihtsustatud, olid kirjutatud täpsed vormistamise reeglid ja tekstide lugemise protsess koos kvantitatiivsete tunnuste arvutamisega on automatiseeritud juba R koodi abil.

Tulevaseks ennustamiseks tudengite esseede põhjal arvutati ja lisati andmestikku C järgmised atribuudid:

1. teksti leksikaalse mitmekesisuse näitaja, ingl *lexical diversity* (lexdiv),
2. plagiaadi protsent – dokumendi teiste dokumentidega Urkundi andmebaasis sarnasuse skoor (plagiaat),
3. Flesch-Kincaidi teksti loetavuse hinnang, ingl *readability score* (flesch_kin),
4. „vesise" näitaja – stoppsõnade arvu ja tekstis olevate sõnade koguarvu suhe (vesi),
5. teksti „rämps" – kõige sagedamini esinevate sõnade arvu ja teksti sõnade koguarvu suhe (spam).

Esseede teksti teisenduste ja uute tunnuste detailne kirjeldus koos valemitega asub Lisas 4.

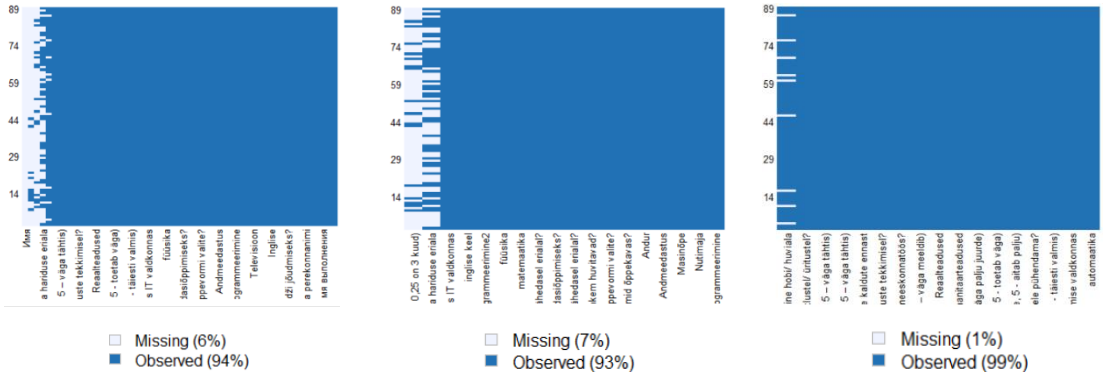
3.3 2020. a. abiturieentide küsitlus (andmestik D)

2020. aastal koostas telemaatika ja arukate süsteemide eriala töörühm ankeedi eesti ja vene keeles (vt Lisa 5) ning pakkus seda täita kolledžisse IKT-erialale sisseastumisel. 2020. aastal täitis küsimustiku vabatahtlikult 46 sisseastujat. Ankeetküsitluse kasutamiseks eemaldati uuringus nende üliõpilaste vastused, kes ei läbinud vestlust või ei kinnitanud oma soovi kolledžis õppida. Pärast neid vähendamisi muutus objektide arv 46-st 30-ks. Andmestikku lisati veerg esimese õppeaasta väljalangemise andmetega.

Elektroonilise ankeedi eesmärk oli välja selgitada sisseastujate motivatsioon, ettekujutus valitud telemaatika ja arukate süsteemide erialast ja kogemus selles valdkonnas. Ankeet koosneb 31 küsimusest, mis annavad üldteavet tulevasest üliõpilasest, tema nägemusest valitud suunal ja mõningatest isikuomadustest. Koostatud küsimustik oli jaotatud

kolmeks osaks: üldandmed demograafiliste küsimustega, eriala ja kolledži valiku täpsustavad küsimused ja isiklikud sotsiaalsed omadused. Ankeet sisaldab valdavalt struktureeritud küsimusi (85% küsimuste koguarvust). Selliseid küsimustikke on lihtsam töödelda, kuna vastaja valib sobiva vastuse konkreetsest vastuste loendist, mis vähendab trükivigade ja vastusevariantide arvu. Kui struktureeritud küsimus ei eelda mitut või dihhotoomset valikut, mõõdetakse seda intervalliga 1 kuni 5, mida nimetatakse Likerti skaalaks.

Ankeedis oli 4 küsimust, millele vastuste hulgas olid puuduvad väärtused (vt Joonis 6).



Joonis 6. Küsitluse puuduvate andmete visualisatsioon.

Ankeet nõudis eeltöötlust ja puhastamist. Täielik puhastamise ja uute tunnuste loomise protsess on toodud Lisas 6.

4. Metoodika

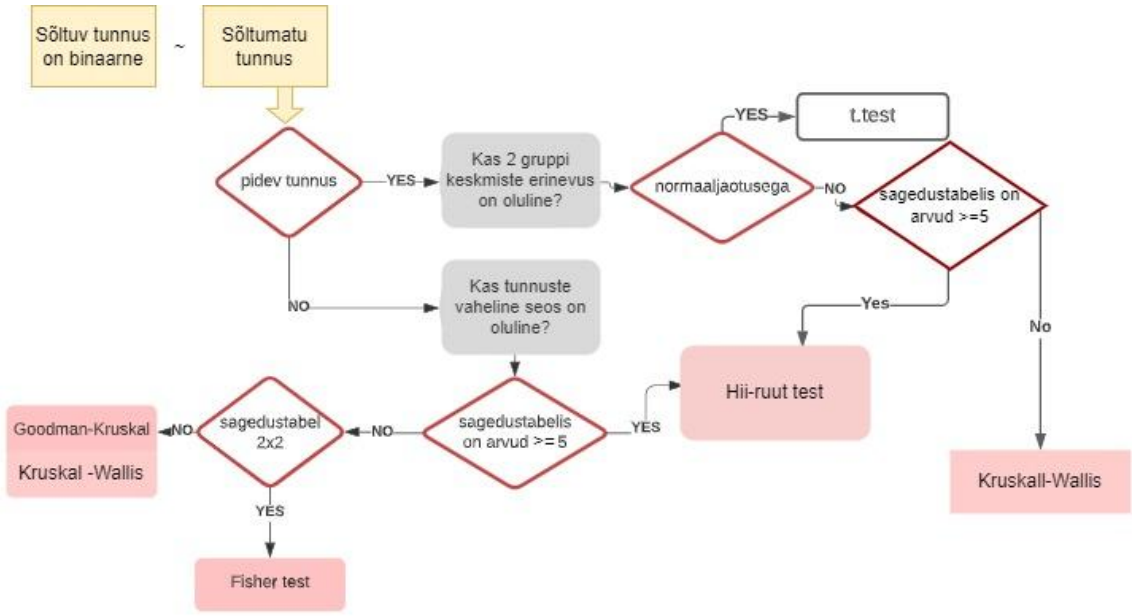
Järgnevalt kirjeldatakse kõiki meetodeid, mida kasutatakse selles töös tunnuste võimalike seoste analüüsimiseks ja mõjutavate tegurite tuvastamiseks, et rakendada neid varajase väljalangemise prognoosimudeli koostamisel.

4.1 Andmeanalüüsi statistilised meetodid

Potentsiaalse seose avamiseks oli kasutatud dispersioon- ja korrelatsioonanalüüsi ning šansside hinnangut. Seost nimetati oluliseks, kui oli tõestatud tunnuse mõju sihttunnusele *y*, väljalangemise indikaatorile, testide abil. Testidel on kaks hüpoteesi: H0 - tunnused on statistiliselt sõltumatud ja H1- tunnused on statistiliselt sõltuvad. Kui testi olulisuse tõenäosus on väiksem kui 0.05 (p-value < 0.05), siis loetakse H1 hüpoteesi tõestatuks ja mõlemate tunnuste vahel on sõltuvus. [10]

Andmetüübid jagunevad kategoorilisteks ja arvulisteks. Kategooriliste andmete uurimine algab nominaalsetest (binaarsetest), järjestikest ja lõpeb intervallidesse grupeeritud andmetega, s.o nominaalsed< järjestikud< intervallidesse grupeeritud. Hierarhia kõrgema taseme tunnuste puhul saab rakendada samu meetodeid, mis madalama taseme puhul, kuid mitte vastupidi [11].

Uuritav tunnus, üliõpilase väljalangemise indikaator, on binaarne, see peab sõltuma teistest andmestiku tunnustest, mis peavad olla üksteisest sõltumatud. Kuna meie andmestike suurused ei ole suured ja erindite olemasolu juhul pole võimalik objekte eemaldamist, siis sõltumatu muutuja mõju tuvastamiseks binaarmuutujale kasutati dispersioonanalüüsi mitteparameetrilisi meetodeid, nagu hii-ruut, Goodman-Kruskal gamma ja tau, Kruskal-Wallise ja Fisheri testid. Skeem Joonisel 7 seletab käesolevas töös testi valimise põhjusi.



Joonis 7. Testide valemise skeem.

Kvalitatiivsete andmete analüüsimeetodeid rakendatakse juhul, kui nii sõltumatu kui sõltuv tunnus on kas nominaalsed või ordinaalsed (järjestusandmed). Sellise analüüsi aluseks on hii-ruut test (χ2-test), mille jaoks, kui uuritav tunnus on binaarne tunnus, andmed esitatakse 2xn sagedustabeli kujul. Hii-ruut testi tulemus on usaldusväärne ainult siis, kui vähemalt 80% unikaalsetest väärtustest peab eeldatav sagedus olema rohkem kui 5 ja ühegi pole sagedusega vähem kui 1 [12]. Antud tingimuste mittekehtivuse või väiksema valimi juhul kasutatakse alternatiivteste, nagu Fisheri [13] ja Goodman-Kruskal tau teste [14] nominaaltunniste või Kruskal-Wallis [15] ja Goodman-Kruskal gamma [16] teste järjestatud tunnuste jaoks, viimane sobib hästi küsitluse viiepunktilise Likerti tüüpi vastuste analüüsiks [17].

Fisheri (täpne) test on hii-ruut testi alternatiiv 2x2 sagedustabeli ja ka väikeste valemite puhul. Kuna hii-ruut testi p-väärtus on arvutatud ligikaudseid meetodeid kasutades ja aproksimatsioon ei ole adekvaatne hii-ruut testi eelduste mittekehtimise juhul, siis täpne Fisheri test on täpselt arvutatud [18].

Pideva tunnuse binaarse tunnuse klasside keskmiste erinevuse kindlakstegemise jaoks kasutatakse Kruskal-Wallise testi [19], mis ei nõua pideva tunnuse normaaljaotuse tingimust. Kui selle testi olulisuse tõenäosus on väiksem kui 0.05, siis on rühmade vahel olulisi erinevusi.

Šanss on uuritava sündmuse toimumise ja mittetoimumise tõenäosuste suhe. Šansside suhe (ingl *odds ratio*; OR) näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss vaadeldavas rühmas võrreldes kontrollrühmaga, [20, lk 10]. Šansside suhe on üks binaarse tunnuse kvantitatiivsetest hinnangutest, mis väljendab selle mõju uuritavale binaarsele tunnusele [21, lk 311]. Enne kasutamist olid kõik pidevad tunnused jaotud klassideks ja edasi teisendatud andmestiku iga tunnus dihhotoomseks kujuks ehk kahe väärtustega kas 0 või 1. Kahemõõtmelise sagedustabeli puhul šansside suhe võib arvutada valemi järgi või R keele funktsiooni *fisher.test*() abil, mis tagastab ka šansside suhte arvu [20].

Kuna dihhotoomsed tunnused on kvantitatiivsed tunnused, siis viimasena oli kasutatud korrelatsioonanalüüsi ja Pearsoni korrelatsioonikordajat [22, ptk 2.6].

Mudeli võimalikud probleemid võivad tekkida multikollineaarsuse pärast ehk seletavate tunnuste vahel tugeva seose pärast, siis sel etapil kontrollitakse lineaarset sõltuvust tunnuste vahel. Töös oli andmetest loodud uued tunnused ja võib ilmuda multikollineaarsus, mille põhjuseks on sama nähtuse kirjeldamine erinevate tunnustega. Multikollineaarsuse selgitamiseks kasutati korrelatsioonimaatriksit, mille kordajate suured väärtused saavad olla multikollineaarsuse põhjuseks. Korrelatsioonanalüüs koos logistilise mudeli VIF arvutusega[6] (ingl *variance inflation factor*) oli kasutatud multikollineaarsuse avastamise jaoks. Leitud üldistatud dispersiooni mõju faktor, mis on suurem 10, näitab tugevat multikollineaarsust [22, ptk 10.7].

4.2 Masinõppe meetodid

Töö põhiliseks eesmärgiks on esmakursuslaste varjase väljalangemise prognoos. Sihtmuutuja *y*, väljalangemise indikaator, on binaarne tunnus väärtusega 1 (tudeng on eksmatrikuleeritud esimese õppeaasta jooksul) või 0 (tudeng ei ole eksmatrikuleeritud esimese õppeaasta jooksul). Kuna probleemiks on binaarne klassifitseerimine, siis lahendamiseks on mudel, mis suudaks võimalikult täpselt kaht klassi eraldada.

[6] Selles töös oli kasutatud R keele paketi *car* funktsiooni VIF(), mille kood on esitatud Lisas 8.

4.2.1 Klassifitseerimise algoritmid

Klassifitseerimise mudeli ehitamiseks oli kasutatud R keele paketti *caret* ja selle funktsiooni *train*(), mis töötab paljude nii regressiooni kui klassifitseerimise algoritmidega. Selles töös oli kasutatud otsustuspuu (rpart), AdaBoosti (Ada), Naive Bayesi (NB), tugivektor-masina (SVM), osavähimruutude diskriminantanalüüsi (PLS-DA) ja logistilise regressiooni (LogR) meetodeid.

- SVM: tugivektor-masina algoritmi eesmärk on leida andmetest hüpertasandi, mis annab suurima eraldust andmeeksemplaride vahel. SVM töötab ainult kahte klassi klassifitseerimiseks. Algoritm ei ole tundlik andmekogumi suuruse ja äärmuslikute väärtuste suhtes [23, lk 293]. Selles töös kasutati *svmRadial* (Support Vector Machines with Radial Basis Function Kernel) radiaalse tuumafunktsiooni [21, lk 384] andmete mittelineaarse eraldamise jaoks. Meetodi puuduseks on enam kui kahe argumendiga mudelite raske tõlgendamine.
- NB: Naive Bayesi algoritm, mis põhineb Bayesi teoreemil ja loetakse statistiliseks klassifitseerimismeetodiks. Selle meetodi eeliseks on vastupidavus andmete müra vastu ja suutlikkus hinnata parameetreid väikese treeningkomplekti põhjal [24].
- Rpart: otsustuspuu on binaarne klassifikatsioonipuu. Puustruktuuris terminalsõlmedes ehk „lehtedes“ määratakse ennustatav klass, otsustussõlmedes (alates juursõlmest) on esindatud jagunemise protsessi tingimused. Puu interpriteerimine on lihtne, kirjeldatakse operaatorite *if-then* komplektina. Meetod on tundlik andmete muutmise suhtes ja ei ole optimaalse ennustuse efektiivsusega [21, lk 203]. Selles uuringus valitakse klassifikatsioonipuu juurutamiseks algoritm Rpart.
- LogR: logistiline regression hindab seletavate tunnuste mõju sihttunnuse ühe klassi esinemise tõenäosusele šansi logaritmi kaudu. Meetodi eeliseks on lihtsus järelduste tegemises mudelis lineaarkombinatsioonina esitatud argumentide mõju kohta [21, lk 320]. Meetod on tundlik tunnuste vahelise tugeva kollineaarsuse suhtes.
- Ada: AdaBoost on nõrkade klassifikaatorite adaptiivne võimendus (*boosting*). Algoritm genereerib nõrkade klassifikaatorite jada ja iga iteratsiooni korral leiab parim nendest andmete kaalude põhjal [21, lk 429]. AdaBoost töötab binaarseks

klassifitseerimiseks. AdaBoost on tundlik andmete kvaliteedi ja müra suhtes. AdaBoosti eeliseks on mudelite kõrge ennustusvõime.

- PLS-DA: osavähimruutude diskriminantanalüüs. Osavähimruutude meetod (PLS) on universaalne meetod nii pideva kui kategoriaalse tunnuse ennustamiseks. PLS meetodil klassifikatsiooni nimetatakse PLS-DA. Algoritm eraldab kaht klassi kasutades rühmade vahelise informatsiooni ja põhinedes PLS meetodi kovariatsiooni maksimeerimise printsiibile. Meetodit soovitati kasutada juhtudel, kus ei saa garanteerida objektide suurt arvu (vähemalt 5-10 korda suurem tunnuste arvust). [21, lk 330-333]. Meetod ei sobi iga andmete kogumile.

Mudelite ehitamiseks kasutatud R keele paketi *caret* funktsioon *train*() lubab valitud algoritmi parameetrite häälestamist ja valib „optimaalse“ mudeli põhinedes mudeli tulemuslikkuse hinnangutele [25, ptk 5]. Mudeli tulemuslikkuse hinnangud on kirjeldatud järgmises alapeatükis. Joonisel 8 on näidatud selle töö mudelite loomise koodi.

```
train(x=X, #treeninguosa andmed ilma sihttunnuseta
 y=trainClass, #treeninguosa sihttunnuse väärtused
 method="svmRadial, nb, rpart, glm, pls, adaboost", #algoritm
 preProc=c("center","scale"),
#enne algoritmi rakendamist oli kasutatud andmete teisendused nagu
#keskendamine ja skaleerimine, arvude suuruse mõju algoritmile SVM, LogR, PLS
#eemaldamiseks
        metric="AUC",
#tagastatakse kahe klassi klassifitseerimise mudeli tulemuslikkuse hinnangud.
 trControl=fitControl, #kontrollmeetodid
 tuneLength = 10 #algoritmi parametrite analüüs (ei tööta glm jaoks)
)
```

Joonis 8. Mudeli loomise kood.

Kasutatud kood tagab iga algoritmi mudeli loomise ühtlase struktuuri. Andmete teisendust kasutatakse tegurite väärtuste ühisel skaalal muutmise jaoks, kus tsentreerimine („*center*“) tähendab teguri väärtustest tema keskväärtuse lahutamist ja normeerimine („*scale*“) — dispersiooniga jagamist.

4.2.2 Tulemuste valideerimine ja hinnangud

Kuna uuringu andmeid ei ole piisavalt palju, siis andmete treening- ja testimisandmeteks jagamise asemel kasutati mudeli ehitamisel 10 kordusega 5-kordset

ristvalideerimist (5-*fold cross validation*). Andmeid jaotakse 5 osaks, selline jagamine on võrdne andmestiku treening- ja testimisandmeteks jagamisega suhtes 4:1. Mudel kontrollitakse 4 korda, kusjuures iga osavalimit kasutatakse treeninguandmetena parajasti üks kord. Viimasel sammul arvutatakse mudeli hinnangu keskmine näitaja. Ristvalideerimise meetod vähendab ülesobitamise (*overfitting*) efekti, kuid *k*-kordne ristvalideerimine suurendab väikese valimi korral tulemuslikkuse hinnangu hajuvust. Peale 5-kordset ristvalideerimist kasutati ka jäta-üks-vahele ristvalideerimist (*Leave-One-Out Cross Validation*), mida soovitatakse kasutada väikestel või tasakaalustamata andmetel mudeli hindamisel [26]. Jäta-üks-vahele ristvalideerimine korratakse *n* korda ja igas järgmises iteratsioonis mudeli treenimisel osaleb *n* – 1 objekti.

Katsete tulemuste võrdlemiseks ja mudeli efektivsuse kontrollimiseks olid valitud veamaatriksi (*Confusion Matrix*) põhjal arvutatud järgmised mõõdikud:

- F1 skoor (*F1 Score*) [27], mis võtab arvesse nii täpsust (*precision*) kui saagist (*recall*) ja võrdub nende harmoonilise keskmisega. Antud hinnang hästi töötab tasakaalustamata klasside juhul, näidates kuidas hästi on ennustatud väiksem positiivne klass.
- Kappa kordaja (*Kappa*) normeerib klasside leitud kokkulangevust klasside juhuslikul paiknemisel oodatava kokkulangevusega [28]. Kappa kordaja interpreteerib mudeli ennustusvõime järgmisel skaalal, mis pakus Cohen [29]: < 0.2 vilets; < 0,4 madal; < 0.6 keskmine; < 0.8 hea ja < 1 väga hea. Kappa mõõdiku eeliseks on selle väärtuste skaala, mis lubab erinevaid klassifikaatore omavahel võrrelda. Coheni Kappat soovitatakse kasutada tasakaalustamata andmetel.

Töös modelleerimise etapp oli realiseeritud kahel ristvalideerimise meetodil (Joonis 9).

```
fitControl <- trainControl(
 method = "repeatedcv, LOOCV",
# mudeli korduv ristvalideerimine või jäta-üks-vahele ristvalideerimine
 number = 5,  # mitmeks osaks jagada valimu (valikul)
 repeats = 10,# valideerimise korduste arv
 summaryFunction=prSummary,# F1 skoori ja Kappa kordaja tagastamiseks
 classProbs=TRUE,# kas arvutada klasside tõenäosusi koos ennustusväärtustega
 savePredictions = TRUE
#salvestab ennustusi optimaalsete häälestusparameetrite korral
 )
```

Joonis 9. Mudelite valideerimise ja hindamise kood.

Selle töö üks püstitatud ülesannetest on leida andmed, mille kaudu võib tõsta ennustusvõimet Weka programmis leitud mudeli ennustustäpsusega 70%. Kuna Weka mudeli korral SAIS'ist saadud üliõpilaste andmed olid tasakaalustatud, siis sel juhul F1 skoor on ennustustäpsusega (*Accuracy*) lähedane ja edasi sobib mudelite võrdlemiseks.

Tunnuste mõjukus hinnatakse funktsiooni *varImp* (*Importance variables for models*) abil paketist *caret*. Funktsiooni sisend on mudel ja funktsioon tagastab vastavalt kasutatud algoritmile [25, ptk 15] mudeli tähtsamaid tunnuseid kahenevas järjekorras ja tulemusi võib graafiliselt esitada.

4.2.3 Tasakaalustamata ja väikeste andmete probleemid

Käesoleva töö andmestike sihttunnuse klasside tasakaalustamatus on kirjeldatud Tabelis 1 ja andmestike ühendite tasakaalustamatus on esitatud Tabelis 2.

Tabel 2. Töös kasutatud andmestike ühendite tasakaalustamatus.

Andmestik	**Andmete kirjeldus**	**Andmete aastad**	**Objekti-de arv**	**y = 1 katkestas esimese õppeaasta jooksul**	**y = 0 ei ole katkesta-nud**	**väljalangemise %**
B+C	Üliõpilaste andmed enne õppimise algust SAIS'ist + Vestluse andmed + Esseede kvantitatiivsed andmed	2019 ja 2020	60	19	41	32% (tasakaalustamata)
B+C+D	Üliõpilaste andmed enne õppimise algust SAIS'ist + Vestluse andmed + Esseede kvantitatiivsed andmed + Küsitluse andmed	2020	22	7	15	32% (tasakaalustamata)

Tabelitest on selge, et andmestikes D ja B+C+D on kogutud vähe andmeid. Andmeanalüüsis kehtib printsiip, mida rohkem infot sisaldab valim, seda väiksem võiks

olla valimi suurus ja vastupidi [30], siis andmeanalüüsiks antud andemete arv on piisav. Vaid kogutud andmeid ei ole piisav masinõppe algoritmide kasutamiseks.

Tasakaalustamata andmekogumitel ehitatud masinõppe mudelitel on madal võime ennustada vähemusklassi objekte. Masinõppe algoritmid ei tööta üldjuhul tõhusalt tasakaalustamata ja väike objektide arvuga andmestike korral. Töös on rakendatud algoritmid, mis on soovitatud kasutamiseks väikestel andmetel. Töötades väikeste andmetega oli märkinud, et (1) algoritmi üldine jõudlus sõltub andmete jaotusest ja mitte suurusest ja (2) piiratud andmetel töötavad kõige usaldusväärsem algoritmid (jõudluse kahenemise järgi): AdaBoost, Naive Bayes, SVM, juhumets (*Random Forest*) ja neurovõrgud [31]. Veel oli leitud, et andmemahu suurendamine 100st objektist 1000-ni mõjub positiivselt mudeli ennustustäpsusele ja väheneb umbes 20% võrra RMSE vea suurust [32]. Väike andmete arv raskendab mustri tuvastamist, kuid kasutatakse teatud kindlusega seaduspärasuste kohta eelduste jaoks. Valimi suurus on seotud mudeli täpsusega negatiivselt, see tähendab et mudelid annavad ülepaisutatud tulemused väikestel andmetel. Selline probleem lahendatakse ristvalideerimise abil, millest kirjutati punktis 4.2.2.

Klasside tasakaalustamatus mõjub väiksema klassi ennustusele. Sel juhul soovitatakse erinevaid meetodeid väiksemate valimite juhul nagu mudeli häälestamine trahvikaalude kaudu või uute andmete sünteesimine [21]. Kuna klasside tasakaalustamatus loetakse problemaatiliseks kui vähemusklassi osakaal on alla 10% ja selle töö andmete tasakaalustamatus loetakse kergemaks, siis autor ei kasuta mainitud meetodeid.

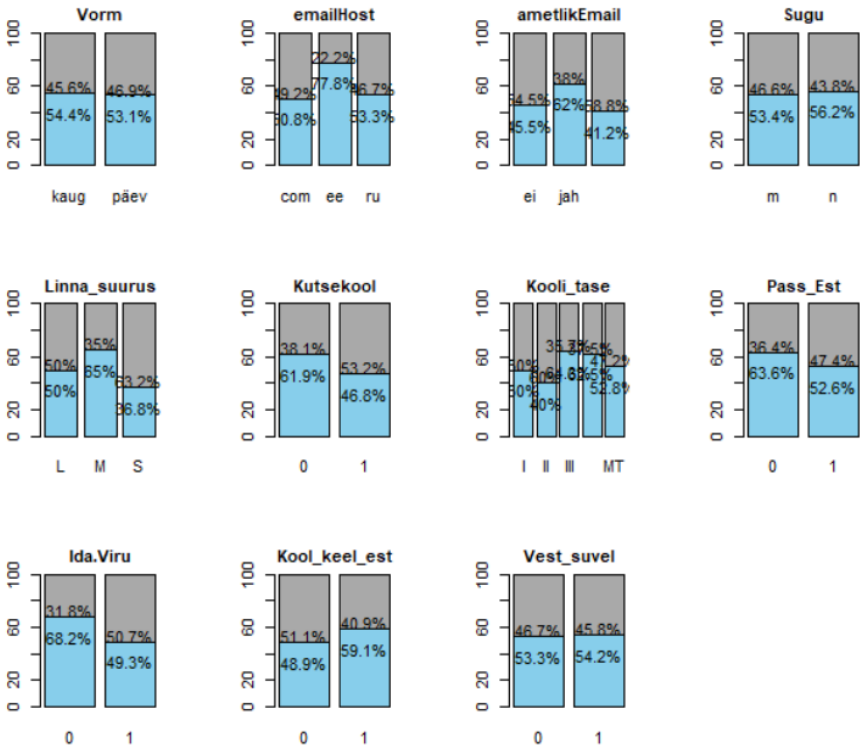
Otsustuspuu ei ole tundlik tasakaalustamata andmete suhtes [33], SVM soovitatakse raske tasakaalustamatuse juhul [34], AdaBoost kasutatakse laialdaselt tasakaalustamata andmetel [35].

5. Andmeanalüüsi meetodite rakendamine ja saadud tulemused

Andmeanalüüsi eesmärgiks on muutujate tuvastamine, mis on statistiliselt olulised sõltuva muutuja seletamiseks. Iga andmestiku kohta esitas autor eeldusi, mida selles osas kontrolliti.

5.1 Andmestiku B analüüs ja tulemused

Andmestik sisaldab 89 esmakursuslase andmeid ja pärast teisendust omab 17 tunnust kokku. Joonisel 10 tumehalli värviga on määratud positiivse klassi (eksmatrikuleeritud) ja helesinisega – negatiivse klassi objektide osakaal kategooriliste tunnuste rühmades.

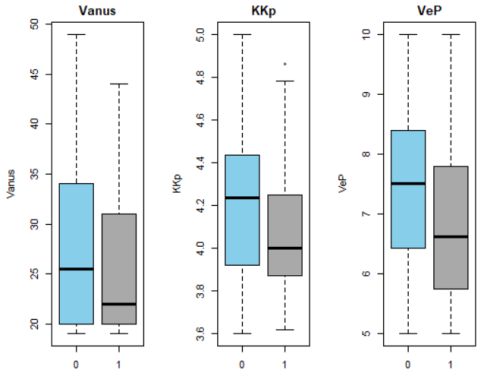


Joonis 10. Andmestiku B sihttunnuse kategooriliste tunnuste rühmadesse jaotus (tumehall värv on positiivne klass „eksmatrikuleeritud“).

Jooniselt on näha, kuidas sihttunnuse klasside jaotus erineb sõltuvalt vaadeldatud tunnustest, näiteks väljalangemine on üle 50% väikelinnadest pärit, Ida-Virumaal

elavate või kutsekooli lõpetanud tudengitel. Erinevuste olulisus kontrollitakse testide kaudu, mille tulemused on toodud järgmises alapeatükis.

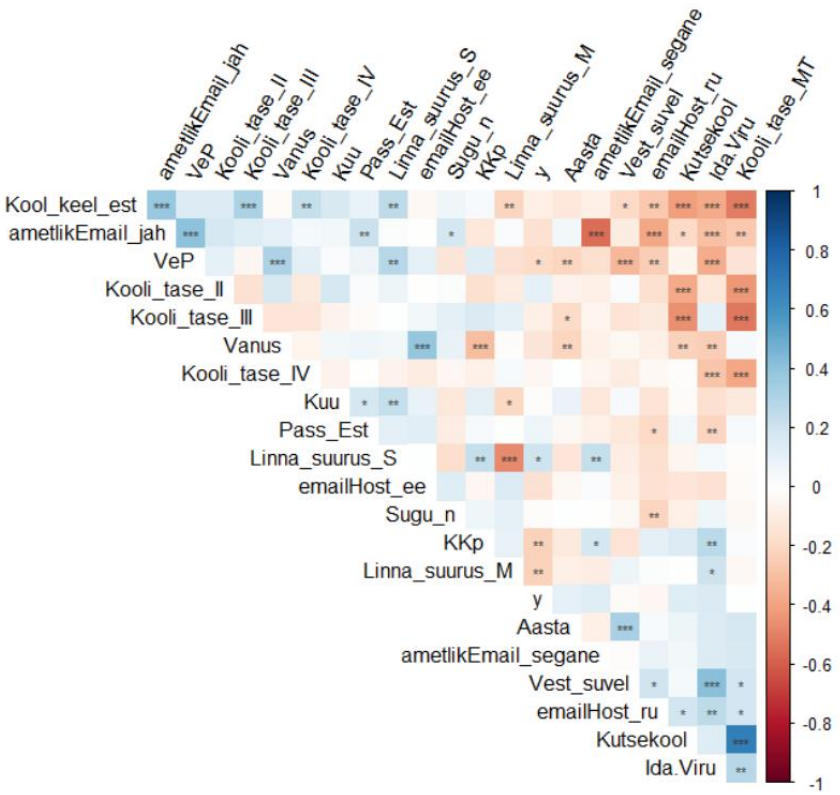
Joonisel 11 karp-vurru diagrammid illustreerivad arvuliste tunnuste keskmiste erinevusi sihttunnuse rühmades.



Joonis 11. Andmestiku B kvantitatiivsete tunnuste keskmiste erinevus sihttunnuse rühmades.

Inimesed, kes katkestasid oma õppimist oma I aasta jooksul, on noorem ning keskmiselt madalama keskhariduse lõputunnistuse keskmise hinde (*Kkp* $\bar{x}_{y=1} = 4.06$) ja vestluse keskmise tulemusega (*VeP* $\bar{x}_{y=1} = 6.7$).

Korrelatsiooni maatriksi (Joonis 12) olulised korrelatsioonikordajad (p-value < 0.1*, <0.05**, <0.01***) näitavad negatiivset lineaarset sõltuvust keskmise linna suuruse (*Linna_suurus_M*), keskhariduse lõputunnistuse keskmise hinde (*Kkp*), vestluse punktide (*VeP*) ja sihttunnuse vahel ning positiivset sõltuvust väike linna suuruse (*Linna_suurus_S*) ja sihttunnuse vahel. Korrelatsioonimaatriksi arvutamisel kvalitatiivsete tunnuste baasil olid konstrueeritud binaarsed tunnused.



Joonis 12. Andmestiku B korrelatsioonimaatriks.

Arvuliste tunnuste vahel jälgitakse kõige suurem positiivne seos (r = 0.69) kutsekooli ja „määramata“ kooli tasemega (kui koolinimetus puudub riigieksamite keskmiste tulemuste pingereas) ja suurem negatiivne seos *Kooli_tase_*III ja *Kooli_tase_MT* vahel (r = -0.52). Logistilise regressiooni VIF kontroll näitas kõige suurema tugevusega multikollineaarsust tunnusel Kooli_tase (GVIF = 20.89).

5.1.1 Testide tulemused

Tunnuste mõju sihttunnusele oli kontrollitud erinevate testide abil. Testide valik vastas Joonisel 7 esitatud skeemile. Näiteks t-testi[7] abil oli tõestatud keskhariduse lõputunnistuse (*KKh*) keskmise hinde keskmiste erinevuse olulisus sihttunnuse rühmades ja ei olnud tõestanud, et vestluse tulemus (*VeP*) mõjutab väljalangemist oluliselt. Ülejäänud testide tulemused on esitatud Tabelis 3.

Tabel 3. Andmestiku B testide tulemused

Test	**Tunnus**	**p-value/ kordaja**	**Mõju tugevus**
t-test (Walch)	KKh	0.0274	väike (Cohen’s d)

[7] VeP ja KKp tunnused on normaajaotusega

Test	Tunnus	p-value/ kordaja	Mõju tugevus
Pearsoni korrelatsioonikordaja	KKh	-0.23	negatiivne nõrk
	Linna_suurus_M	-0.23	negatiivne nõrk
	Linna_suurus_S	0.19	positiivne väga nõrk
	VeP	-0.18	negatiivne väga nõrk
Hii- ruut	Ida-Virumaa	0.0385	mõõdukas mõju (Crameri V)
Fisheri test	KKh_3.7_4.1	0.0161	OddsRatio = 3.02
	VeP_<6	0.0263	OddsRatio = 3.19

Järeldus: punktid vestluse eest avaldavad väga nõrka negatiivset mõju väljalangemisele. Üliõpilased, kes said vestluse tulemuse, mis on ligilähedane minimaalsele vastuvõetavale piirmäärale, jätavad 3.19 korda suurema tõenäosusega kolledži pooleli.

Multikollineaarsuse kontroll leidis potentsiaalse probleemi tunnusel *Kooli_tase*.

5.1.2 Eelduste kontroll

Selle andmestiku puhul oli püstitud ja kontrollitud mõned eeldused, mille tulemused on esitatud allpool Tabelis 4.

Tabel 4. Püstitatud eeldused ja kontrolli tulemused.

Eeldus	Kontrolli tulemus
1. Naiste keskmine koolihinne erineb meeste keskmisest koolihindest. 2. Keskmine koolihinne sõltub lõpetatud haridusasutusest (kas kutsekool või mitte). 3. Meeste vestluse keskmine tulemus (7.25) erineb naiste keskmisest (6.8). 4. Õpilase keskmine koolihinne mõjutab vestluse tulemust.	Ei olnud tõestatud kontrolli käigus.
5. Sisseastujate vanus mõjub vestluse tulemusele. 6. Missugusest maakonnast on sisseastuja mõjub vestluse tulemusele. 7. Matemaatika riigieksami tulemus sõltub lõpetatud keskhariduse õppeasutusest	On tõestatud: p-value < 0.01***, Pearsoni korrelatsioonikordaja 0.30. On tõestatud: Kruskall-Wallis testi p-value=0.041. On leitud keskmise mõjuga assotsiatsioon: Goodman-Kruskall tau test, xy = 0.54.

5.1.3 Andmestiku B eelanalüüsi järeldused

Andmestiku B eelanalüüsist autor on teinud järgmised järeldused: 2019. ja 2020. aastate sisseastujad on keskmise vanusega 27 aastat, enamus juhtudel on meessoost, Eesti kodakondsusega ja Ida-Virumaalt, kes lõpetas kas gümnaasiumi või kutsekooli, mis ei asu esimestes kohtades riigieksamite pingireas, lõputunnistuse keskmise hindega 4.15, kes tuli vestlusele suvel ja said keskmiselt 7.2 punkti.

Antud analüüs andis esimesed oletused andmestiku tunnuste ja sihttunnuse, väljalangemise indikaatori, seose kohta. On leitud, et kogutud sisseastujate andmed nagu sugu, valitud õppevorm, rahvus, vanus ei anna kasulikku infot, sest pole tõestatud testide abil nende mõju väljalangemisele ja see tulemus langeb teiste uuringute järeldustega kokku.

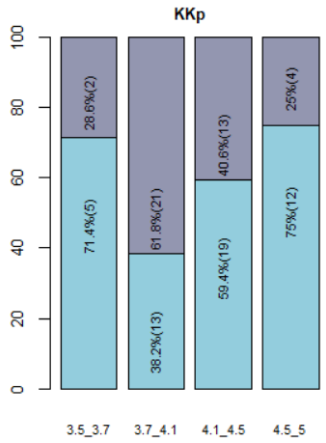
Kuna kolledži sisseastumisel arvestatakse keskhariduse lõputunnistuse keskmine hinne (*Kkh*) ja sisseastumise vestlusel saadud punktide arv (*VeP*), autor analüüsis neid tunnusi lähedasem. On leitud, et tudengitel, kellel Kkh = [3.7;4.1), väljalangemise šanss suureneb 3 korda (OddsRatio=3.02). Koolitunnistuse keskmise hinde üldine mõju kolledžist väljalangevusele on nõrk ja negatiivne. Korrelatsioonanalüüs näitab, et *Kkh* on suurem Ida-Virumaalt (nõrk ja positiivne seos) või väikestest linnadest sisseastujatel (nõrk ja positiivne seos) ja on väiksem vanematel sisseastujatel (tunnusega vanus on nõrk ja negatiivne seos). Antud seoseid võib põhjendada selline fakt, et Ida-Virumaalt tulevad enamasti õppima Jõhvi, Sillamäe või Narva kutsehariduskeskuste noored lõpetajad, kelle koolitunnistuste keskmised on üldiselt head, millest vestluskomisjon on teadlik ja sellepärast puudub seos *Kkh* ja *VeP* vahel (vt Joonis 12).

Vestlus mõjub väga nõrgalt väljalangevusele ja ainult liiga madala vestluse tulemusega (*VeP*<6) on 3.2 korda suurem šanss loobuda õppimisest (OddsRatio=3.19). Vestluse punktidele mõjub negatiivselt sisseastuja maakond (madalam on neil, kes on pärit Ida-Virumaalt), suvine läbiviimise aeg (suvel on madalam) ja positiivselt mõjub vanus (vanematel sisseastujatel on kõrgem).

Neid kaks tunnust, *Kkh* ja *VeP*, millistel põhinevad kolledži sisseastumistingimused, nõrgalt mõjuvad esmakursuslaste väljalangevusele, kuid oli leitud, et sisseastujate seas, kelle vestluse tulemus on keskmisest kõrgem ($VeP\ \bar{x} = 6.7$), väljalangemise protsent on 28.2%. Aga neil, kelle keskmine tulemus on üldisest keskmisest madalam, on 39.5%.

Arvestades eelpool tehtud järeldused vestluse tulemuse kohta pakub autor vestluse lävendi kuni 6 punktini tõsta.

Kolledžis planeeritud keskhariduse lõputunnistuse keskmise hinde tõstmises kuni 3.8 ei näe autor suurt kasu, sest kutsehariduskeskuse sisseastujate lõputunnistuse keskmine hinne on 4.192, minimaalne 3.62. Sel hetkel ainult 7.9% sisseastujatel lõputunnistuse keskmine on alla 3.7 ning Joonis 13 näitab suurima väljalangemise protsenti Kkh väärtuste lõigul 3.7 – 4.1.



Joonis 13. Keskhariduse lõputunnistuse keskmise hindega lõigul 3.7 - 4.1 esmakurslaste seas on kõige suurem väljalangemise protsent (värvitud tumedama värviga).

Kuna ainult tunnusel *Kooli_tase* VIF on suurem kui 10, siis multikollineaarsus on tugev ja edaspidi autor ei kasuta antud tunnust mudelite ehitamisel. Korrelatsioonimaatriksist on selge, et tunnus sõltub Kutsekool tunnusest, sest kutsekooli lõpetajad väga harva sooritavad riigieksameid ja kutsekoolide eksamite tulemused puuduvad pingereas.

5.2 Esseede andmestiku C analüüs ja tulemused

Esseed kirjutatakse sügissemestri lõpus ja semestri jooksul toimub üliõpilaste õpingutest loobumise loomulik protsess ja selle tõttu esseede arv vähenes ja autor sai analüüsida ainult 67 esseed. Kolledžis õpivad tudengid vene ja eesti emakeelega. Kolledži õppekeel on riigikeel (eesti keel), kuna kolledž on tolerantne teiste keelte suhtes on lubatud kirjutada esseed üliõpilasele mugavas keeles. Edaspidi olid tõlgitud kõik esseede tekstid inglise keelde, et täielikult kasutada R keele tekstianalüüsi paketi *quanteda* (*Quantitative Analysis of Textual Data*) kõiki võimalusi. Esseede andmestikku C oli lisatud andmestiku B tunnused nagu sugu, vanus, õppevorm, keel (vt Joonis 14).

Variable	Stats / Values	Freqs (% of Valid)	Graph
esseKeel [factor]	1. est 2. rus	41 (61.2%) 26 (38.8%)	
aasta [factor]	1. 2019 2. 2020	32 (47.8%) 35 (52.2%)	
sugu [factor]	1. m 2. n	53 (79.1%) 14 (20.9%)	
õppevorm [factor]	1. päeva 2. sessioon	26 (38.8%) 41 (61.2%)	
y [factor]	1. 0 2. 1	46 (68.7%) 21 (31.3%)	

Joonis 14. Esseede andmestiku C kirjeldus.

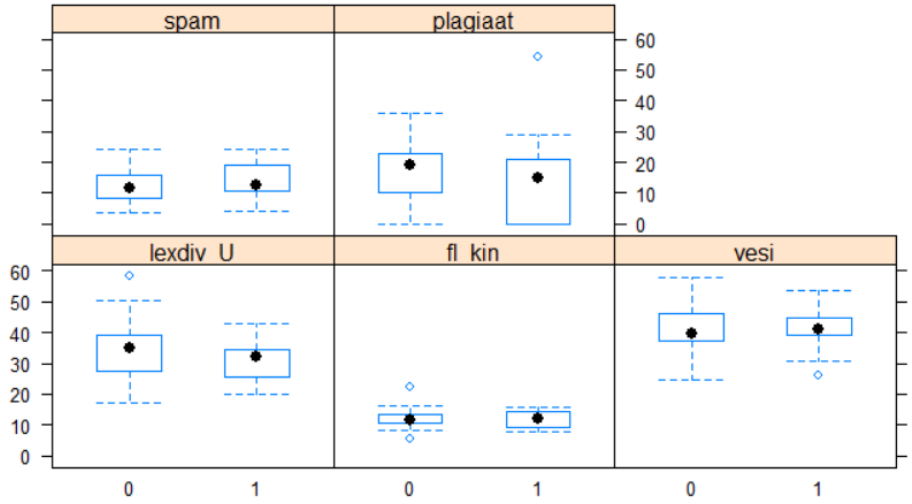
Keskmine essee koosneb 332 sõnast ja 14 lausetest (vt Joonis 15).

Types	Tokens	Sentences
Min. : 36.0	Min. : 44.0	Min. : 2.00
1st Qu.:136.0	1st Qu.:252.5	1st Qu.:10.00
Median :169.0	Median :324.0	Median :14.00
Mean :165.3	Mean :332.6	Mean :14.31
3rd Qu.:200.5	3rd Qu.:396.0	3rd Qu.:17.50
Max. :271.0	Max. :638.0	Max. :29.00

Joonis 15. Esseede tekstide arvkarakteristikud.

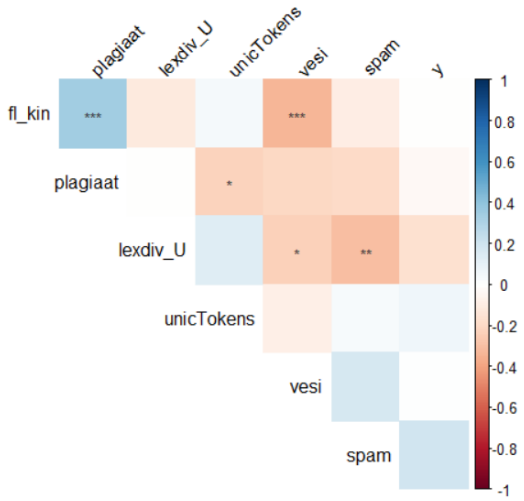
Autori poolt olid arvutatud andmestikku C teksti iseloomustavad näitajad nagu teksti leksikaalne mitmekesisus, plagiaadi protsent, loetavuse hinnang, teksti „vesise" ja „rämpsu" näitajad.

Vaadates leksikaalse mitmekesisuse erinevust sihttunnuse rühmades (vt Joonis 16) oli leitud, et õppimist katkestanud tudengite esseede leksikaalne rikkus on keskmiselt madalam ($\bar{x}_{y=1} = 31.18$ ja $\bar{x}_{y=0} = 34.00$). Esseede loetavuse hinnang oli leitud Flesch-Kincaid valemi kaudu, kus hindamisaste võrdub USA hariduseastmega ja näitab teksti mõistmiseks vajalikku haridusetaset. Esseede loetavuse keskmine tase $\bar{x}_{y=1} = 11.95$ ja $\bar{x}_{y=0} = 11.93$, mis vastab 11. klassi haridustasemele.



Joonis 16. Esseede teksti iseloomustavate näitajate keskmiste erinevus sihttunnuse rühmades, kus lexdiv_U – leksikaalne mitmekesisus, fl_kin – teksti loetavus.

Korrelatsioonmaatriks (vt Joonis 17) näitab kõige suuremat positiivset seost teksti loetavuse (*fl_kin*) ja plagiaadi vahel (r = 0.34) ning negatiivset seost teksti loetavuse ja „*vesise*“ vahel (r = -0.33).

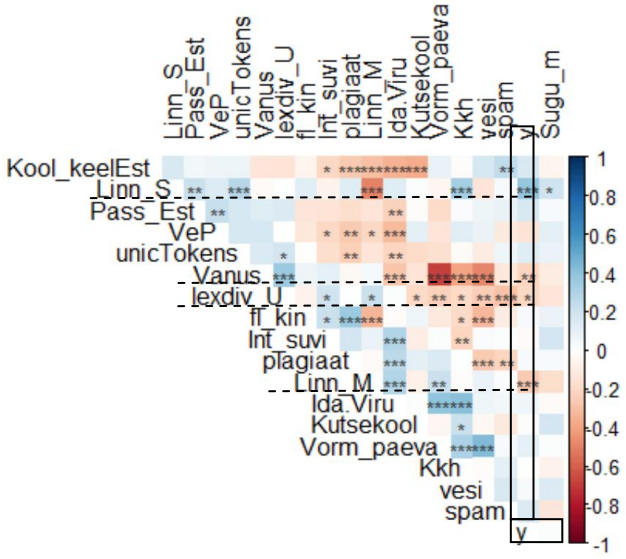


Joonis 17. Andmestiku C korrelatsioonimaatriks.

Sel etapil ei olnud leitud olulist seost sihttunnuse *y* ja esseede kvantitatiivsete tunnuste vahel mitte ühe statistilise meetodi abil. Logistilise regressiooni mudeli tunnuste VIF < 1.4 ning multikollineaarsus puudub.

5.3 Vestluse ja esseede ühise andmestiku B+C analüüs ja tulemused

Andmestike ühendamine suurendab tunnuste arvu, mis rikastab informatsiooni. Vestluse ja esseede andmestikute ühendamisel objektide arv väheneb kuni 60. Selle ühise andmestiku puhul on kõige tugevama negatiivse lineaarse seosega *Vanus* ja *Vorm_paev* (r = -0.69), *vesi* ja *Vanus* (r = -0.49). Kõige tugevama positiivse lineaarse seosega on *Linn_S* ja *Linn_M* (r = 0.5), *vesi* ja *vorm_paeva* (r = 0.43) (vt Joonis 18).



Joonis 18. Andmestikute B+C korrelatsioonimaatriks.

Sellest korrelatsioonimaatriksist on näha ka neli nõrka olulist korrelatsiooni sihtmuutuja ja atribuutide *Linn_S* (r = 0.39), *Linn_M* (r = -0.26), *Vanus* (r = -0.22) ja *lexdiv*_U (r = -0.21) vahel. Meenutame, et enne oli leitud sihtmuutuja oluline seos tunnustega KKh (eriti KKp_3.7_4.1), Linna_suurus_M, Linna_suurus_S, VeP (eriti VeP<6), Ida-Virumaa.

Samal ajal näitasid mitme muutujaga logistilise regressiooni mudeli VIF väärtused vahemikus 1.62–5.83 tugeva multikollineaarsuse puudumist.

5.4 Küsitluse andmestiku D analüüs ja tulemused

Küsitlus koosneb 31 küsimustest (vt. Lisa 4). Aastate 2020-2021 jooksul 89 sisseastuja vastasid küsimustele. Selles töös on esitatud ainult 2020 aasta vastuseid (30 inimest 46-st alustas oma õppimist kolledžis), sest edasi ühendatakse andmestik D andmestikutega B ja C, et vaadata koos teiste tunnustega küsitluse vastuste mõju väljalangemisele.

Küsitluse küsimustes, kus on vaja eriala tundmist ning õpinguteks ettevalmistust ja isiklikke oskusi hinnata, oli kasutatud Likerti skaala 1-st 5-ni. Hindamiskaala punkt 3 on neutraalne paik. Kui vastuste keskmine on alla 3 või üle 3, siis selle vastuse jaotus on asümmeetriline. Vastuste vasakpoolne asümmeetria (vt Lisa 7) räägib suurema hindade 4 ja 5 osakaalust või positiivsest hindamisest ja parempoolne asümmeetria – negatiivsest hindamisest. Küsimuste analüüs on jagatud neljaks osaks, et vastata asjahuvilistele teemadele nagu keelte oskus, erialast informatiivsus, sisseastujate eelteadmised ja sotsiaalsed oskused.

Sisseastujad väga kõrgelt hindavad oma vene keele oskust ($\bar{x}$~4.24) ja madalam eesti ($\bar{x}$~3.87) ja inglise keele oskust ($\bar{x}$~3.52). 58.7% mitte eesti keele emakeelega sisseastujatelt hindavad oma eesti keele oskust nagu suurepärane ja hea, kusjuures eesti keele oskuse keskmine enesehinnang muu emakeelega sisseastujatel on 3.8 ja vene emakeelega on 3.5 vastavalt. Sisseastujate geograafia ei ole väga lai (vt Joonis 19).

	Harju maakond	Ida-Viru maakond	Lääne-Viru maakond	Rapla maakond	Tartu maakond	Viljandi maakond
eesti	42.1	16.4	66.7	100.0	100.0	100.0
muu	10.5	4.9	0.0	0.0	0.0	0.0
vene	47.4	78.7	33.3	0.0	0.0	0.0

Joonis 19. 2020. aasta küsitluse vastajate emakeelte jaotus elukoha maakonna järgi.

Vaadates sisseastujate jaotust maakondade järgi oli leitud, et eesti keele keskmine oskus Lääne-Virumaal on $\bar{x}$~5, Harjumaal – $\bar{x}$~3.9 ja Ida-Virumaal – $\bar{x}$~3.4. Osal sisseastujatel inglise ja vene keele oskus on madal või puudub, kusjuures vene ja inglise keelte oskuste vahel on negatiivne lineaarne seos (vene keele oskus on kõrgem nendel, kelle inglise keele oskus on madalam).

Telemaatika erialast informeeritus oli keskmiselt 3.51, pool sisseastujatelt ei teadnud midagi (hinne 2) või teadsid vähe (hinne 3) oma tuleviku erialast. Joonisel 20 on esitatud sisseastujate keskmised hinded küsimusele „Millega on telemaatika seotud?“ Sisseastujate arvamusel Telemaatika ja arukate süsteemide eriala on kõige rohkem seotud side (K7_Side), programmeerimise (K7_Prog) ja nutimajaga ja kõige vähem meedia (K7_Meed) ja televisiooniga (K7_TV).

Joonis 20. Küsimuse „Millega on telemaatika seotud?“ vastuste diagramm.

Selgus, et sisseastujad positsioneerivad enda tehnikahuvilisena (K22_Reaal $\bar{x} = 3.94$) ja hindavad ennast objektiivselt (82.6%), kuid hindavad oma eelteadmisi matemaatikas, füüsikas ja programmeerimises keskmiselt vastavalt $\bar{x}$~{3.57; 3.25; 3.01}. Eelteadmisi automaatikas on hinnatud madalam ($\bar{x}$ =2.4) ja paljudel puudub praktiline kogemus automatiseerimise valdkonnas ($\bar{x}$ =2.2). Telemaatika eriala on esimene valik umbes 80% sisseastujatel. 48% vastajatest varem haridus oli seotud valitud erialaga, kusjuures kõike vastajate eelhariduse tase on kutsekool. 2020. aasta sisseastuja tunneb huvi programmeerimise (50%) ja arukate süsteemide (20%) vastu.

Sotsiaalsed oskused, mis olid mõõdetud küsimuste „Kas teile meeldib töötada meeskonnas?“ (K23_mskt), „Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega?“ (K27_Suht) ja „Kui tähtis on Teile suhtlemine sotsiaalvõrkudes?“ (K28_SotV) kaudu, on kas positiivse asümmeetriaga või alluvad normaaljaotusele.

„Kui palju Teie arvates aitab kolledžidiplom soovitud töökoha saamisel?“ on hinnatud keskmise hinnanguga 4.49. Üldjuhul kolledžist küsimustele vastused on väga positiivsed.

5.4.1 Testide tulemused

Tunnuste mõju sihttunnusele oli kontrollitud erinevate testide abil. Testide tulemused on esitatud Tabelis 5.

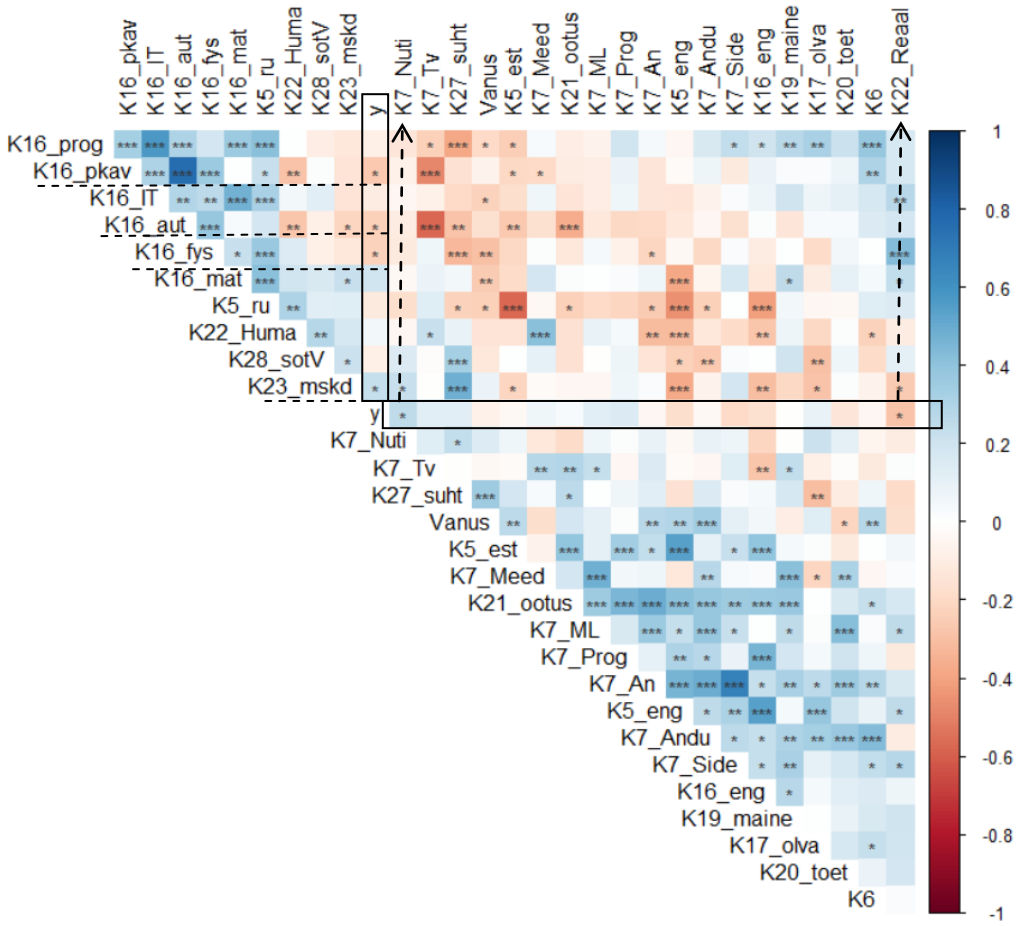
Tabel 5. 2020. a küsimustiku testide tulemused.

Test	Tunnus	p-value/ kordaja	Mõju tugevus
Pearsoni korrelatsioonikordaja	K23_mskd	0.48	positiivne mõõdukas

Test	Tunnus	p-value/ kordaja	Mõju tugevus
Pearsoni korrelatsioonikordaja	K7_Nuti	0.26	positiivne nõrk
	K16_prog	-0.33	negatiivne nõrk
	K16_füüs	-0.32	negatiivne nõrk
	K16_aut	-0.29	negatiivne nõrk
Hii- ruut	Sugu	0.0445	mõõdukas mõju (Crameri V)
	Aeg_kol_<30min	0.0240	mõõdukas mõju (Crameri V)
	Aeg_kol_>1h	0.0265	mõõdukas mõju (Crameri V)
Kruskall-Wallis	Sugu	0.03	suhteliselt tugev mõju (epsilon)
Fisheri test	Sugu_n	0.0472	OddsRatio = 9.42
	Aeg_kol_<30min	0.0227	OddsRatio = 6.90
	Aeg_kol_>1h	0.0292	OddsRatio = 0
Goodman-Kruskali gamma	K7_ML	0.560	suhteliselt tugev
	K14_TooK	1	väga tugev
	K23_mskd	-0.550	negatiivne suhteliselt tugev
	K27_suht	-0.60	negatiivne suhteliselt tugev

Korrelatsioonimaatriks (vt Joonis 21) näitab, et arvuliste tunnuste vahel on 6 tugevat seost ($|r| > 0.6$):

- K16_aut ja K16_pkav (kogemus automaatikast ja praktiline kogemus automaatika valdkonnas) vahel r = 0.87;
- K7_TV ja K16_aut (Kas telemaatika on seotud televisiooniga ja automaatika eelteadmiste hindamine) vahel r = -0.65;
- K5_est ja K5_eng (eesti keele ja inglise keele oskuste hindamine) vahel r = 0.65;
- K5_eng ja K16_eng (inglise keele oskuse hindamine ja eeloskuse hindamine) vahel r = -0.64;
- K16_IT ja K16_Prog (Hinnake oma eelteadmisi IT-s ja programmeerimises) vahel r = 0.63;
- K5_rus ja K5_eng (vene ja inglise keelde oskuste hindamine) vahel r = -0.61.



Joonis 21. Andmestiku D korrelatsioonimaatriks.

Logistilise regressiooni VIF kontroll (vt Joonis 22) näitas multikollineaarsust paljude tunnuste vahel ehk vahetlikute küsimuste olemasolu.

```
    K5_est     K5_eng      K5_ru         K6      K7_Tv    K7_Meed      K7_An    K7_Prog    K7_Nuti      K7_ML
150.784370 156.251916  66.702566   5.585146 236.055526  92.080025  24.385063  69.627626  20.833729   6.507181
  K7_Side    K7_Andu    K16_mat    K16_fys   K16_prog    K16_eng     K16_IT    K16_aut   K16_pkav   K17_olva
 28.914545  14.493513  12.294640 237.704985 170.705466 113.660711  26.409093 109.978470 177.292641  60.707309
 K19_maine   K20_toet  K21_ootus   K22_Huma  K22_Reaal   K23_mskd   K27_suht   K28_sotV
 22.973423  24.865192  45.455964  98.223259 101.265535 121.964957  93.110999  73.604022
```

Joonis 22. Arvuliste küsimuste VIF kontroll.

Üksteisega asendatavate tunnuste eemaldamise pärast jäid 12 järgmist arvulist tunnust: K5_est, K6, K7_ML, K7_Nuti, K7_Andu, K7_Side, K16_aut, K16_mat, K16_prog, K16_eng, K22_Reaal, K23_mskd.

Statistilise analüüsi kaudu leitud sihttunnusele mõjukad tunnused ei ole korreleeritud omavahel ja VIF on intervaalis 1.22 (K7_ML) -5.112 (K14_TööKogemus).

5.4.2 Eelduste kontroll

Samuti leiti, et puudub sõltuvus

- humanitaarse mõttelaadi ja õpingute jätkamise soovi vahel; rühmad on homogeensed (Fligner-Killeen testi p-value=0.1302) ja rühmade vahel puudub erinevus (Kruskal-Wallis p-value=0.4296);
- reaalseteks õppeaineteks kalduvuse ja õpingute katkestamise vahel (Fligner-Killeen testi p-value=0.7610) ja rühmade vahel puudub erinevus (Kruskal-Wallis rank testi p-value=0.1050);
- reaalseteks õppeaineteks kalduvuse ja matemaatika oskuse vahel (Goodman-Kruskal tau xy = 0.1, Goodman-Kruskal gamma testi p-value = 0.2008);
- vanuse ja inglise keele oskuse vahel (Goodman-Kruskal tau xy = 0.051);
- (kodu)tee ja kolledži vahelise kauguse ning õppetundide arvu vahel päevas (Goodman-Kruskal tau xy = 0.14 ja Goodman-Kruskal gamma testi p-value = 0.058, Kruskal-Wallis testi p-value = 0.0954);
- selle vahel, kas üliõpilase elukoht on Ida-Virumaal, ja selle vahel, et üliõpilased peavad diplomi saamist oluliseks faktoriks hea töö saamisel (Goodman-Kruskaltau xy = 0.03, Goodman-Kruskal gamma testi p-value = 0.57);
- selle vahel, kas üliõpilase elukoht on Ida-Virumaal, ja selle vahel, et üliõpilane jätkab õpinguid (Kruskal-Wallis testi p-value = 0.2638);
- üliõpilase emakeele ja õpingute katkestamise vahel (Fisher testi p-value = 0.4661). Aga esimese semestri jooksul on olemas seos eesti keele oskuse ja õpingute katkestamise vahel (r = -0.26).

Statistiliselt kinnitati, et

- eelmiste erialaste kogemuste olemasolu mõjutab automaatikaalaseid teadmisi (Goodman-Kruskall tau yx = 0.328 ja Goodman-Kruskall gamma testi p-value = 0.008 ja gamma=0.744, Kruskall-Wallis testi p-value = 0.0496);
- eesti keele oskus mõjutab inglise keele oskust ja mõju on positiivne keskmine r = 0.65;
- vene keele oskus mõjutab inglise keele oskust ja mõju on negatiivne keskmine r = -0.61;

- reaalseteks õppeaineteks kalduvuse ja füüsika oskuse vahel (Goodman-Kruskall tau yx = 0.33, Goodman-Kruskall gamma testi p-value = 0.0115 ja gamma = 0.64).

5.4.3 Andmestiku D eelanalüüsist kokkuvõte

Küsimustiku edasise kasutamise jaoks pakub autor vähendada küsimuste arvu vahetlikute ja kolledžist küsimuste kõrvaldamise teel.

Küsimustik vastab oma eesmärgile osaliselt, aitab paremini iseloomustada kolledži sisseastujate kogemusi ja eriala tundmist. Küsimustik ei kajasta sisseastujate motivatsiooni.

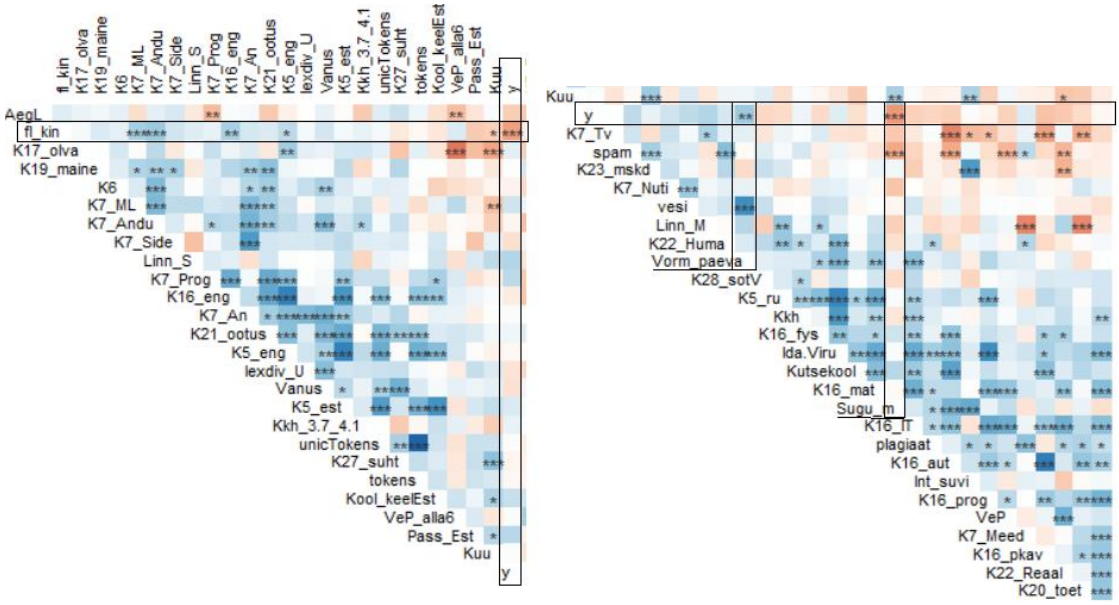
Esimene probleem on eesti keele valdamise tase mitte eesti keele emakeelega sisseastujatel. Kuna võrreldes eesti õppekeelega põhikooli lõpetajatega läheb suhteliselt suur osakaal vene õppekeelega põhikooli lõpetajatest 37-38% kutsekooli, kusjuures valdav enamik õpib vene õppekeelega kutseõppes [36], siis tekib eesti keeles õppimise valmisolekuga probleem. Tudengitele on raske õppida, nad tunnevad ebamugavust ja erijuhtudel dipressiooni ja katkestavad oma õppimist. Meie kolledži Telemaatika eriala jaoks selline probleem esineb suurel määral ainult esimesel semestril ja oli kinnitatud, et üldjuhul keelte oskused ei mõju õppimise katkestamise otsustele. Analüüs avastas kandidaatide inglise keele oskuste erinevust, mis sõltub emakeelest (eesti emakeelega kandidaadil on kõrgem).

Teiseks probleemiks näeb autor eriala tundmist. Kolledži sisseastujatel on ebaselge ja segane ettekujutus tulevikust erialast ja ei ole piisav informeeritus. Kolmas probleem on eelteadmiste taseme ebapiisavus ja ebaühtlus. Kolledž on Tallinna Tehnikaülikooli osa, kuid tehnilistes ainetes eelteadmised on keskmiselt nõrgad ja sisseastujad vääralt hindavad ennast tehnikahuvilistena.

5.5 Vestluse, esseede ja küsitluse ühise andmestiku B+C+D analüüs

Järgmises etapis ühendati kolledži vastuvõtuuringu andmestik D andmekogumiga C+B. Uurimisküsimus oli: kas kolledži vastuvõtuuringu andmed parandavad õpilaste ülikoolieelsetel andmetel, sisseastumisintervjuudel ja esseedel põhinevate mudelite ennustusvõimet. Selle tulemusena saadi ühine andmestik B+C+D kolledži 2020. aastal

vastuvõetud 22 üliõpilaste kirjega. Korrelatsioonimaatriks mugavamaks lugemiseks on jagatud kaheks osaks (vt Joonis 23).



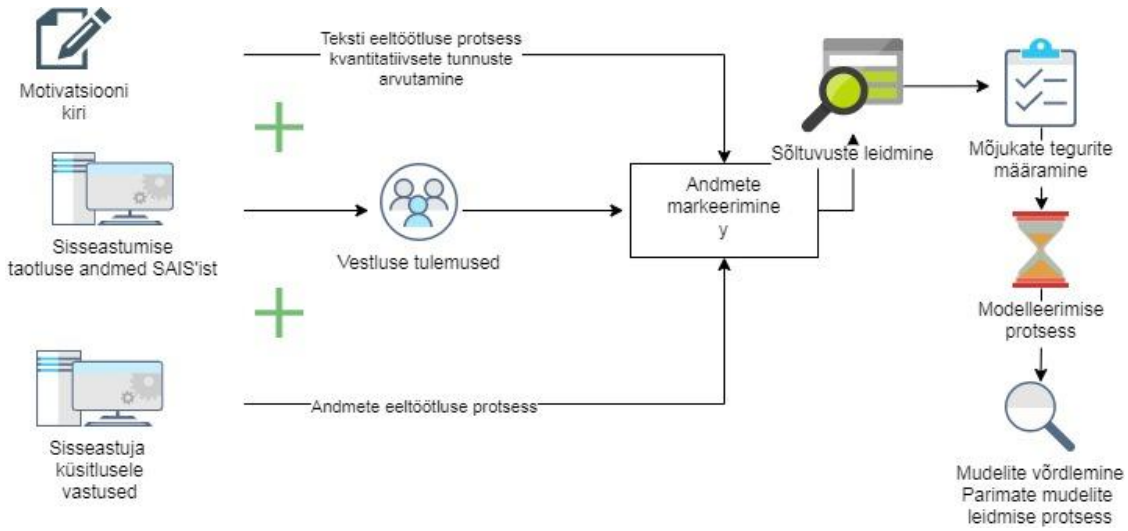
Joonis 23. Andmestike B+C+D korrelatsioonimaatriks.

Sihttunnusele mõjuvad sugu (r = -0.44), teksti loetavus (r = -0.41), plagiaat (r = -0.39) ja päevaõppe vorm (r = 0.37) ning küsitluse küsimustega sihttunnusel olulist seost pole.

Masinõppe rakendamise osas antud andmetel arvestati iga andmestiku kõige mõjukad isikupäratud (nagu sugu, elukoht, emakeel jne) tunnused: VeP_<6, Kkp_3.7_4.1, lexdiv_U, fl_kin, plagiaat, K7_nuti, K16_prog, K16_füüs, K16_aut, K23_mskd, K27_suht. Multikollineaarsuse kontrolli pärast oli võetud plagiaat ja K16_füüs maha.

6. Masinõppe meetodite rakendamine ja saadud tulemused

Joonisel 24 kirjeldatakse antud klassifitseerimisprobleemi protsess, mille tulemusena on andmestikute ja nende ühenduste andmetel leitud parimad ennustusmudelid.



Joonis 24. Tööprotsessi visualisatsioon.

Parima mudeli valimises algpunktiks on 2012-2018 aastate andmetel uuring.

6.1 2012 – 2018 aastate andmetel esimese uuringu parim mudel

Kolledži väljalangemise esmane uuring oli tehtud 2012-2018 aastatel kogutud andmetel, mille klasside jaotus oli tasakaalustatud. Selle uuringu käigus leitud parim mudel ennustustäpsusega 70% oli ehitatud logistilise regressiooni klassifitseerimisalgoritmil ja kasutas ainult kaht tunnust Matemaatika riigieksami punktid, keskhariduse tunnistuse keskmine hinne. Valem (1) esitab mudeli võrrandi [2].

$$y = \frac{1}{1+e^{-(3.42\ -0.02 Mat\_eksam\ -0.73\ Kkh\ )}} \quad (1)$$

Edaspidi loetakse parimaks mudeliks selline konstrueeritud mudel, mille ristvalideerimise keskmine F1 skoor on üle 0.7 ning ristvalideerimise parima mudeli F1

skoor on üle 0.7 ja Kappa kordaja on üle 0.6 (mis vastab kas hea või väga hea ennustusvõimega mudelile).

6.2 2019. ja 2020. aastate andmetel mudelid

Mudeli ehitamisel kasutati andmete eelanalüüsi käigus leitud sihttunnusele mõjutavaid tegureid.

6.2.1 Sisseastujate üld- ja vestluse andmetel mudelid

Tabelis 6 on esitatud SAIS'ist sisseastujate üld- ja vestluse andmetel leitud mudelite hinnangud.

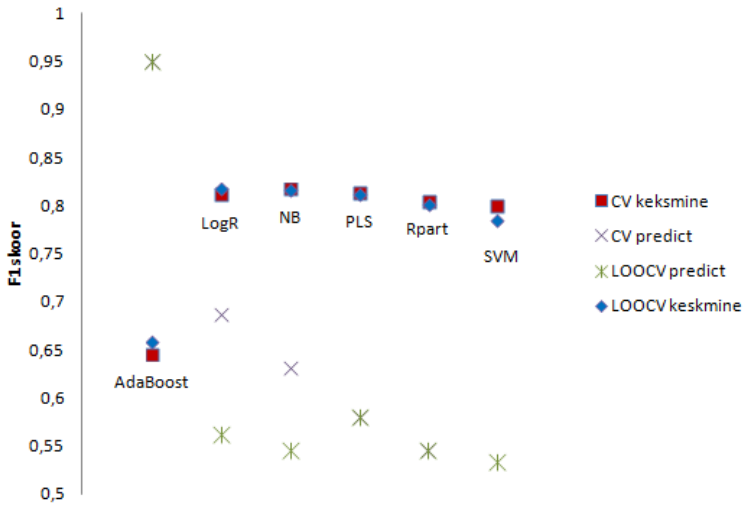
Tabel 6. Andmestikul B ehitatud mudelite tulemused.

B: Sisseastujate üld- ja vestluse andmed

Mõjukad tunnused: Keskhariduse lõputunnistuse keskmine hinne, Linna suurus, Vestluse punktid, kas sisseastuja on Ida-Virumaalt

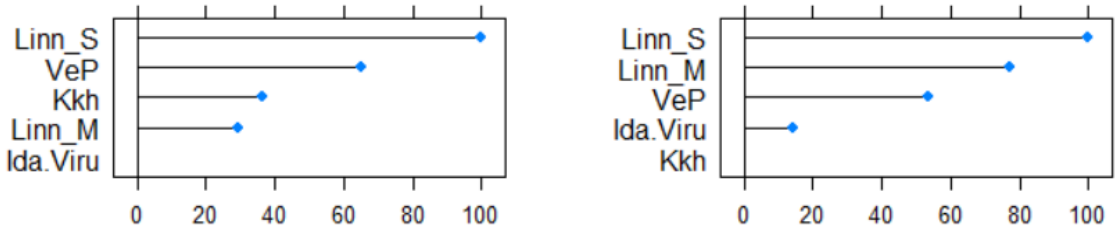
Mudel	**CV**			**LOOCV**		
	F1 kesk	**F1 pred**	**Kappa pred**	**F1 kesk**	**F1 pred**	**Kappa pred**
Ada	0 .6456	0 .95	0 .9251	0 .6597	0 .95	0 .9251
LogR	0 .8118	0 .6878	0 .5576	0 .8182	0 .5625	0 .4109
NB	0 .8172	0 .6316	0 .4608	0 .8177	0 .5455	0 .3785
PLS-DA	0 .8132	0 .5806	0 .4444	0 .8123	0 .5806	0 .4444
Rpart	0 .8048	0 .5455	0 .3785	0 .8017	0 .5455	0 .3785
SVM	0 .8	0 .1	0 .0706	0 .7862	0 .5333	0 .3922

LOOCV metood mõjus tugivektor-masina algoritmi mudelile. F1 skooride väärtuste graafik (Joonis 25) näitab algoritmi *AdaBoost* erinevust teistest algoritmidest.



Joonis 25. Andmestikul B ehitatud mudelite F1 skooride graafik.

Otsuste vastuvõtmisel algoritmid ei baseeru kõikidel tunnustel. Joonisel 26 on algoritmide AdaBoost ja logistilise regressiooni oluliste tunnuste jadad, ülejäänud algoritmidel need pildid korduvad.



Joonis 26. Andmetel B mudelite tähtsad muutujad: (a) logistiline regressioon ja (b) AdaBoost.

6.2.2 Sisseastujate üld-, vestluse ja esseede andmetel mudelid

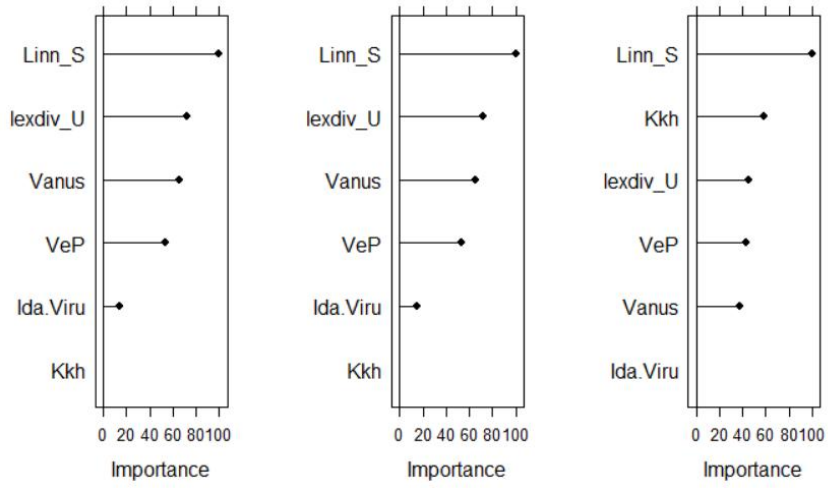
Edaspidi rikastame andmeid esseede kvantitatiivsete teguritega ja kordame katseid. Antud rikastamine muutus kõrgemaks kõikide mudelite jõudlust keskmiselt 22% võrra (vt Tabel 7). Meetodil LOOCV ehitatud AdaBoost mudel on väga hea (Kappa = 1, F1kesk = 0.7671), selle mudeli järgi iga objekt on õigesti klassifitseeritud (F1pred = 1). SVM mudel on hea (Kappa = 0.7928, F1pred = 0.8485, F1kesk = 0.7825).

Tähistame selle mudeli radiaaltuuma parameetreid: cost C = 8, sigma σ = 0.1285.

Tabel 7. Andmestikutel B+C ehitatud mudelite tulemused.

B+C: Sisseastujate üld-, vestluse ja esseede andmed **Mõjukad tunnused:** Keskhariduse lõputunnistuse keskmine hinne, Linna suurus väike, Vestluse punktid, Kas sisseastuja on Ida-Virumaalt, Vanus, Essee leksikaalne rikkus						
Mudel	**CV**			**LOOCV**		
	F1 kesk	**F1 pred**	**Kappa pred**	**F1 kesk**	**F1 pred**	**Kappa pred**
Ada	**0 .7467**	**1**	**1**	**0 .7671**	**1**	**1**
LogR	0 .8059	0 .6857	0 .5576	0 .8193	0 .6857	0 .5576
NB	0 .7702	0 .4848	0 .2956	0 .7838	0 .4848	0 .2956
PLS-DA	0 .8106	0 .6471	0 .5102	0 .8130	0 .6471	0 .5102
Rpart	0 .7656	0 .6829	0 .5197	0 .7214	0 .6829	0 .5197
SVM	**0 .7826**	**0 .8125**	**0 .7475**	**0 .7825**	**0 .8485**	**0 .7928**

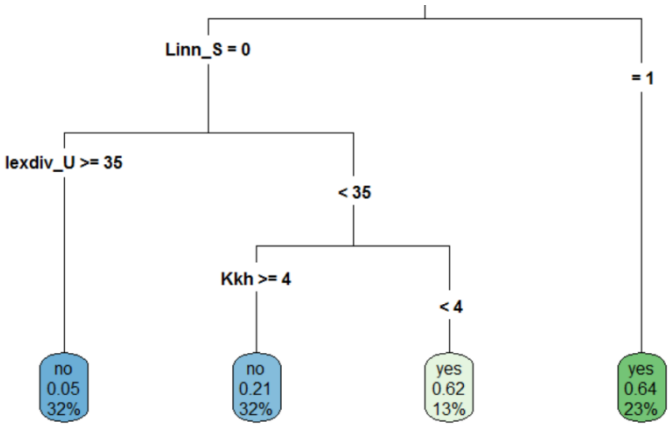
Algoritmid otuste vastuvõtmisel ei arvesta kas keskhariduse lõputunnistuse keskmist või Ida-Virumaal elukohta (vt Joonis 27).



(a) (b) (c)

Joonis 27. Andmetel B+C mudelite tähtsad muutujad: (a) AdaBoost, (b) tugivektor-masin, (c) logistiline regressioon.

Otsustuspuu Rpart mudel ei kasuta vestluse punktid tingimusena, mis annab võimalust kasutada mudelit enne vestluse läbimist.



Joonis 28. Otsustuspuu Rpart.

Joonis 28 näitab, et otsustuspuu algoritmi järgi inimene katkestab õppimist esimese aasta jooksul, kui tuleb liiga väikest (alla 5000 elanikut) elukohast või tema essee leksikaalne rikkus on vähem kui 35 ja keskhariduse lõputunnistuse keskmine on alla 4.

6.2.3 Küsitluse andmetel mudelid

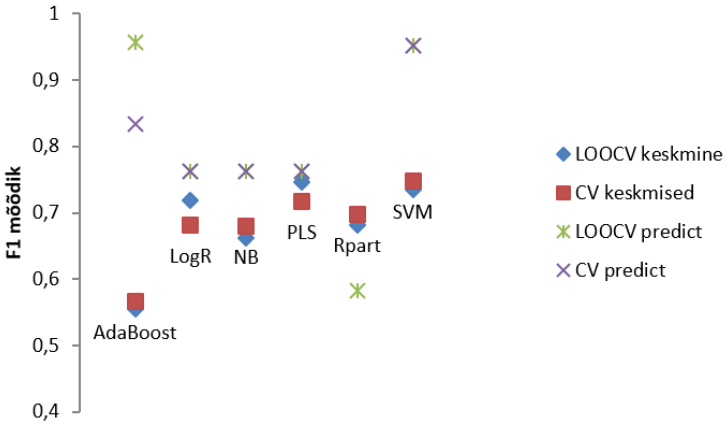
Küsitluse mõjukatel tunnustel ehitatud mudelite tulemused on ülepaisutatud (vt Tabel 8) selle väikse andmestiku jaoks ja masinõppe tulemusi vaatleme nagu seaduspärasuste esmased algeeldused.

Tabel 8. Andmestiku D ehitatud mudelite tulemused.

D: küsitluse andmed **Mõjukad tunnused:** Sugu „naine“; K4: Kui palju aega keskmiselt kulub Teil kolledži jõudmiseks?(alla 30 min) K7: Millega on telemaatika seotud (nutimaja; masinõpe)? K16: Hinnake oma eelteadmisi, mis on vajalikud edasiseks õppimiseks (programmeerimine, füüsika, automaatika)? K14: Kas Teil on töökogemusi telemaatika erialal või sellele lähedasel erialal? K23: Kas Teile meeldib töötada meeskonnas? K27: Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega?

Mudel	**CV**			**LOOCV**		
	F1 kesk	**F1 pred**	**Kappa pred**	**F1 kesk**	**F1 pred**	**Kappa pred**
Ada	0.5670	0.8333	0.7293	0.5548	0.9565	0.931
LogR	0.6816	0.7619	0.6404	**0.7180**	**0.7619**	**0.6404**
NB	0.6790	0.7619	0.6404	0.6627	0.7619	0.6404
PLS-DA	0.7178	0.7619	0.6404	**0.746**	**0.7619**	**0.6404**
Rpart	0.6975	NaN	0	0.6808	0.5833	0.3231
SVM	0.7472	0.9524	0.9281	**0.7354**	**0.9524**	**0.9281**

Tabelis 8 F1 skoor NaN ja Kappa 0 tähendab sihttunnuse positiivse klassi ennustuste puudumist. Sel juhul mudel ei suuda üldse eristada positiivse klassi objekte ja loeb iga objekt negatiivseks. Esitame saadud tulemusi graafikuna (Joonis 29), nendel andmetel peale Rpart algoritmi ülejäänud mudelite F1 skoor ületab 0.7 ja eriti väga hea ennustusvõimega AdaBoost ja SVM mudelid, kuid siin ristvalideerimise keskmiste F1 väärtuste järgi loetakse parimateks algoritmiteks SVM, PLS-DA ja LogR. SVM mudel on väga hea (Kappa = 0.9281, F1kesk = 0.7293, F1pred = 0.9524). Tähistame selle mudeli radiaaltuuma parameetreid: cost C = 8 ja sigma σ = 0.0631.



Joonis 29. Andmestikul D ehitatud mudelite F1 skooride graafik.

Siin, väikestel andmetel, ennustusmudelitel AdaBoost ja SVM on suur erinevus ristvalideerimise ja ennustuste F1 väärtuste vahel.

Mudelite viis esimest olulisemat tunnust on: Kuidas nutimajaga seob sisseastuja telemaatika erialat, Kas on töökogemus IT valdkonnas, Kuidas meeldib meeskonnas töötada, Kas lähedal elab kolledžist, Sugu, Eeloskused automaatikas või füüsikas. Vaadates logistilise regressiooni mudeli ainult statistiliseid tähtsamaid tunnusi {K14_TooK, K23_mskd, Aeg_kol_alla30min, Sugu_n, K16_fys} uue mudeli uuritud F1kesk/F1_pred/Kappa näitajad on 0.7805/ 0.7826/ 0.6548 vastavalt.

Valem (2) esitab selle logistilise regressiooni mudeli valemi, kus kordajad on ümardatud kümnendikeni.

$$y = \frac{1}{1+e^{-(0.3+9.7\cdot Aeg\_kol\_alla15min + 7.5\cdot K23\_mskd + 6.7\cdot K14\_TooK + 5.7\cdot Sugu\_n + 0.3\cdot K16\_fys)}} \qquad (2)$$

6.2.4 Sisseastujate üld-, vestluse, esseede ja küsitluse andmetel mudelid

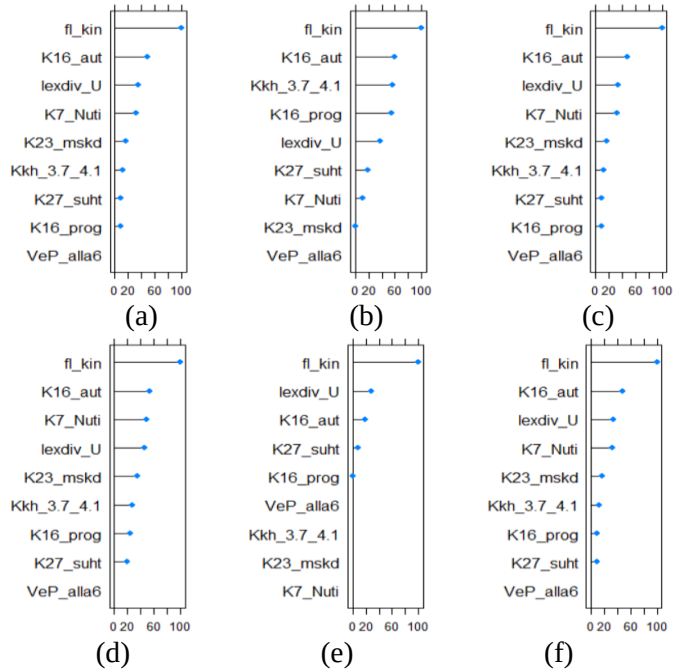
Viimasena oli valmistatud kõikide andmestike ühend objektide arvuga 22. Siin kasutab autor kaht tunnuste loetelu: 1) Tabelis 9 on nimetatud iga andmestiku kõige mõjukamad isikupäratud tunnused ja 2) keskhariduse lõputunnistuse keskmine hinne koos matemaatikas teadmiste enesehinnanguga. Teine loetelu on sarnane esimese uuringu parima mudeli tunnustega, siin autor kontrollib võimalust asendada matemaatika riigieksami tulemus matemaatikas teadmiste enesehinnanguga.

Tabel 9. Andmestikutel B+C+D ehitatud mudelite tulemused.

B+C+D: Sisseastujate üld-, vestluse, esseede ja küsitluse andmed
Mõjukad tunnused: VeP: alla 6 vestluse punkt, Kkp_3.7-4.1: keskhariduse lõputunnistuse keskmine hinne. K7: Millega on telemaatika seotud (nutimaja)? K16: Hinnake oma eelteadmisi, mis on vajalikud edasiseks õppimiseks (programmeerimine, automaatika). K23: Kas teile meeldib töötada meeskonnas? K27: Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega? Essee leksikaalne rikkus, Teksti loetavus.

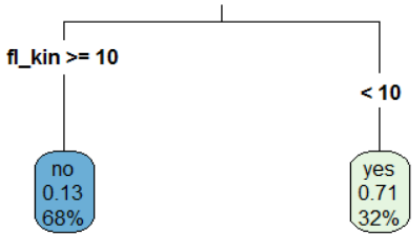
	CV			**LOOCV**		
Mudel	**F1 kesk**	**F1 pred**	**Kappa pred**	**F1 kesk**	**F1 pred**	**Kappa pred**
Ada	0.8357	0.9333	0.8991	**0.7042**	**1**	**1**
LogR	0.5950	0.7692	0.6733	0.6	0.7692	0.6733
NB	0.6734	0.6667	0.4545	0.5270	0.7368	0.56
PLS-DA	0.6530	0.5455	0.4086	0.5754	0.7143	0.581
Rpart	0.8142	NaN	0	0.7974	0.7143	0.581
SVM	0.7874	NaN	0	0.7812	NaN	0.1647

Vaadates mudelite tunnuste olulisust (Joonis 30) saab järeldada, et iga algoritm kasutab tunnuseid: teksti loetavus (*fl_kin*), eelteadmised automaatikas (K16_aut), essee leksikaalne rikkus (*lexdiv*_U) ning üldse ei ole kasutatud vestluse punkte (*VeP*_alla_6).



Joonis 30. Andmetel B+C+D mudelite tähtsad muutujad: (a) adaBoost, (b) logistiline regressioon, (c) Naive Bayes, (d) PLS-DA, (e) ostutuspuu ja (f) tugivektor-masin.

Andmestikute ühendamisel kõikide mudelite tähtsamaks tunnuseks on teksti loetavus (fl_kin). Otsustuspuu koosneb juurtipust funktsiooniga kas *fl_kin* ≥ 10, mis edasi ei käivita jagunemise protsessi aga annab kahe lõpp-punkti „*ei*“ või „*jah*“ (Vt Joonis 31).



Joonis 31. Andmestikutel B+C+D otsustuspuu, mille LOOCV meetodi keskmine F1=0.7975.

Kuna selle otsustuspuu Kappa = 0.581 on kokkulepitud piirist 0.6 väiksem, siis parimaks mudeliks võetakse mudel *AdaBoost* (F1kesk = 0.7042, F1predict = 1 ja Kappa = 1).

Teise katse tunnused on *Kkp*: keskhariduse lõpptunnistuse keskmine hinne ja *K16_mat*: Hinnake oma eelteadmisi matemaatikas. Mudelite tulemused olid väga madalad (vt Tabel 10) ja näitavad, et matemaatika riigieksami tulemust ei saa vahetada matemaatikas teadmiste enesehinnangu vastu.

Tabel 10. Kkp ja matemaatika enesehinnangul mudelite tulemused.

B+C+D: Sisseastujate üld-, vestluse, esseede ja küsitluse andmed

Mõjukad tunnused: *Kkp*: keskhariduse lõputunnistuse keskmine hinne ja *K16_mat*: Hinnake oma eelteadmisi matemaatikas, mis on vajalikud edasiseks õppimiseks

	CV			**LOOCV**		
Mudel	**F1 kesk**	**F1 pred**	**Kappa pred**	**F1 kesk**	**F1 pred**	**Kappa pred**
Ada	0.5754	0.7143	0.485	0.5404	0.8696	0.7887
LogR	0.7262	0.375	0.1892	0.7224	0.375	0.1892
NB	0.6984	0.6667	0.4545	0.6932	0.4286	0.322
PLS-DA	0.7242	0.127	0.3529	0.7114	0.3529	0.127
Rpart	0.6526	NaN	0	0.6152	0.5455	0.2823
SVM	0.7209	NaN	0	0.7417	0.1667	0.1124

Selle peatüki Tabelid 6-9 näitasid, et esimese andmestiku (sisseastujate üldandmed koos vestluse punktidega) mudelite ennustustäpsused ei ole head. Esseede tegurid muutusid ennustusmudelid rikkamaks ja nende lisamisel mudelite ennustustäpsus tõstis keskmiselt 20% võrra. Küsitluse andmed on perspiktiivikad, sest 0.7 piir on kolmel algoritmil ületatud. Kõikide andmete ühendi mudelites ei olnud kasutatud vestluse punkte ning surusid eessee tegurid küsitluse vastused esimeste tähtsamate tunnuste jadast välja.

7. Tulemused

Töö algul oli pakutud 4 hüpoteesi, mille kohta tehakse selles peatükis kokkuvõte.

Sisseastumisvestlus on üliõpilaste väljalangevuse seisukohalt tõhus meede.

Antud hüpotees oli kinnitatud osaliselt töös. Vestluse hindamispunktidel on väga nõrk negatiivne lineaarne seos väljalangemise sihttunnusega ja väiksema vestluse punktiga tudengitel on suurem šanss eksmatrikuleerida esimese õppeaasta jooksul. Vestluse punktid kuuluvad mõjukate tunnuste kogumi ning osalesid klassifitseerimismudelite ehitamises, kuid saadud mudelite ennustusvõimed on keskmised või madalad. Vestluse punkte võtsid arvesse ainult 2 algoritmi (logistiline regressioon ja Naive Bayes) 6-st ja selle tunnuse eemaldamine ei muudanud eriti mudelite täpsuse hinnaguid (vähenesid keskmiselt 10% võrra). Eraldi vestluse punktidel ja keskhariduse lõputunnistuse keskmise hindel ehitatud mudelitest suudsid positiivset klassi eristada ainult Rpart ja AdaBoost algoritmid (F1 skoor 0.37 ja 0.672 vastavalt).

Vestluse punktide panemisel oli märgatud nõrk tendentslikkus. Autor esiteks pakub konkretiseerida hindamissüsteemit, et elimineerida ebaoluliste faktorite mis tahes mõju vestluse tulemusele ja teiseks tõsta vestluse punktide lävendi 6 punktini (tänane on 5).

Küsimustik võimaldab välja selgitada üliõpilaste teadlikkust valitud suunast ja see aitab kaasa prognoosimismudelile.

Antud hüpotees oli kinnitatud töö käigus täielikult. Küsitluse leitud mõjukad tegurid on vastused järgmistele küsimustele: Kui palju aega keskmiselt kulub Teil kolledži jõudmiseks? Millega on telemaatika seotud (nutimaja; masinõpe)? Hinnake oma eelteadmisi, mis on vajalikud edasiseks õppimiseks (programmeerimine, füüsika, automaatika)? Kas Teil on töökogemusi telemaatika erialal või sellele lähedasel erialal? Kas Teile meeldib töötada meeskonnas? Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega? Küsitluse leitud mõjukatel tunnustel täheldatakse viiel algoritmil ennustusvõime hea või väga hea, kuid kolm nendest vastavad parima algoritmi valimise tingimustele: SVM, PLS-DA ja logR. See on antud töö üks kõige

parematest tulemustest ja tõestab, et küsitlus on tähtis klassifitseerimismudelis ja aitab probleemi uurimisel. Autor näeb siin küsitluse andmete arvu suurendamise ja edaspidise analüüsi vajadust praeguste andmete arvu vähesuse tõttu.

Küsitluse analüüsi põhjal (vt Ptk 4.5.3) saab kinnitada, et küsitlus võimaldab paljastada sisseastuja teadlikkust valitud Telemaatika ja arukate süsteemide erialast, kuid autor soovitab vähendada või muuta küsimusi nende tugeva multikkolineaarsuse tõttu. Küsimustik ei kajasta sisseastujate motivatsioonist.

Motivatsioonikirja lisamine sisseastumisel suurendab üliõpilaste väljalangemise ennustamise tõhusust.

Antud hüpotees oli tõestatud. Ühelpool oli tõestatud, et tekstidest arvutatud tegurid ei mõju sihttunnusele, kuid kombineerides teistest allikatest andmetega nende roll tõuseb. Esseede tegurid muutsid ennustusmudelid rikkamaks, nt tunnustel *Keskhariduse lõputunnistuse keskmine hinne*, *Linna suurus*, *Vestluse punktid*, kas sisseastuja on pärit *Ida-Virumaalt* ehitatud mudelite F1pred skooride keskmine on 0.62 ja esseede tegurite lisamisel mudelite F1pred skooride keskmine tõusis kuni 0.724. Töö kinnitas, et teksti arvulised tegurid on tähtsamad kui küsitluse ja vestluse tulemused kõikide andmete ühendis.

Jah, motivatsioonikirja lisamine sisseastumiseprotsessi muutub väljalangemise ennustamise efektiivsemaks. Sel tööl analüüsiti esseede, kuid arvutatud kvantitatiivsed tegurid võib arvutada iga teksti liiki jaoks.

Otsuse tegemiseks piisab eelnevalt kogutud täielikust teabest (*üldandmed*, *vestlus*, *ankeet*, *essee*).

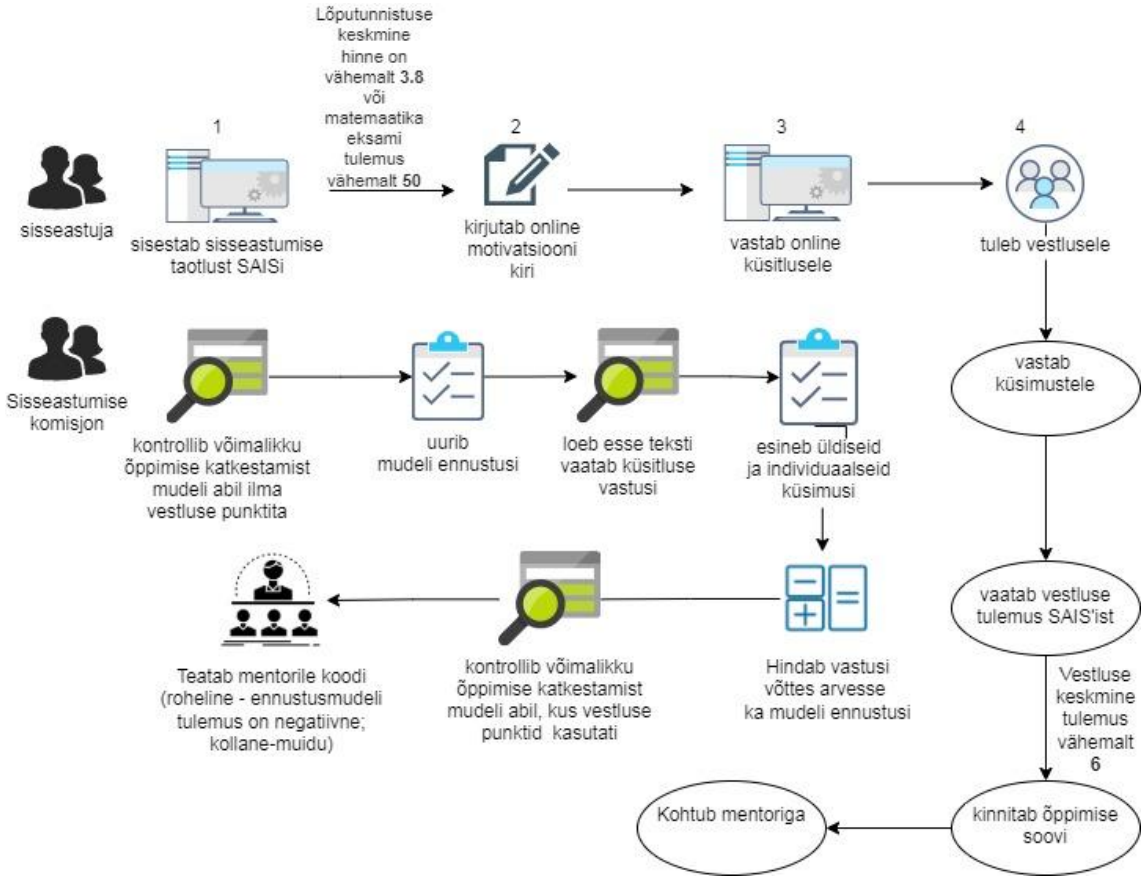
Hüpotees ei olnud kinnitatud täielikult käesoleva töö andmete vähesuse tõttu. Töös on leitud mudelid, mille ennustusvõimed ja F1 skoorid on väga head ja ületavad ettenähtud 70% piiri. Need mudelid, mis vastavad parima mudeli valimise kriteeriumitele, on kogutud Tabelis 11.

Tabel 11. Käesoleva töö masinõppe osa parimad klassifitseerimisemudelid.

Andmestik	**Tunnused**	**Parim mudel**	**F1 kesk/F1 pred/Kappa**
Sisseastujate andmed + esseed (B+C)	Keskhariduse lõputunnistuse keskmine hinne, Linna suurus väike, Vestluse punktid, kas sisseastuja on pärit Ida-Virumaalt, Vanus ja Essee leksikaalne rikkus	(1) AdaBoost	0.7671/1.00/ 1.00
		(2) SVM	0.7825/0.8485/ 0.7928
Küsitlus D	Sugu „naine“; K4: Kui palju aega keskmiselt kulub Teil kolledži jõudmiseks?(alla 30 min) K7: Millega on telemaatika seotud (nutimaja; masinõpe)? K16: Hinnake oma eelteadmisi (programmeerimine, füüsika, automaatika)? K14: Kas Teil on töökogemusi telemaatika erialal või sellele lähedasel erialal? K23: Kas Teile meeldib töötada meeskonnas? K27: Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega?	(3) SVM	0.7293/0.9524/ 0.9281
		(4) PLS-DA	0.7464/0.7619/ 0.6404
	Sugu „naine“; K4: Kui palju aega keskmiselt kulub Teil kolledži jõudmiseks?(alla 30 min) K16: Hinnake oma eelteadmisi füüsikas? K14: Kas Teil on töökogemusi telemaatika erialal või sellele lähedasel erialal? K23: Kas Teile meeldib töötada meeskonnas?	(5) LogR	0.7805/ 0.7826/ 0.6548
Sisseastujate andmed + esseed + küsitlus (B+C+D)	VeP: alla 6 vestluse punkt; Kkp_3.7-4.1: keskhariduse lõpptunnistuse keskmine hinne; K7: Millega on telemaatika seotud (nutimaja)? K16: Hinnake oma eelteadmisi, mis on vajalikud edasiseks õppimiseks (programmeerimine, automaatika)?; K23: Kas Teile meeldib töötada meeskonnas? K27: Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega? lexdiv_U: essee leksikaalne rikkus; fl_kin: teksti loetavus.	(6) AdaBoost	0.7042/1.00/ 1.00

Autori arvamusel kogutud küsitluse andmeid ei ole piisav kokkuvõtte tegemiseks.

Küsitluse analüüs leidis ka, et tudengite eelteadmised ei ole ühtlased ja omavad puudusi. Siin võib lisada ka matemaatika- või komplekstesti sooritamist sisseastumiseprotsessi, mis lubab võrrelda eelteadmiste erinevust. Vestlus on kasulik osa ja olleks efektiivsem, kui seda teha küsitluse ja essee järel, et korrigeerida vestluse küsimusi küsitluse ja essee abil. Siin autor pakkub vaadata vestluse hindamisesüsteemi üle, sest vestluse keskmine hinne ei tööta piisavalt hästi ja kõikide andmete ühendi mudelites ei olnud kasutatud. Kuna vestluse hinne koosneb neljast osast: huvi eriala vastu, tulevikuväljavaated, õpivalmidus, eelnev haridus ja töökogemus, siis oleks hea lisada neid mudelisse eraldi. Joonisel 32 on esitatud võimalik sisseastumisprotsessi järjekord.



Joonis 32. Vastuvõttu võimalik protsess.

8. Kokkuvõte

Käesoleva magistritöö eesmärgiks oli masinõppe meetodite abil luua mudelid TalTech Virumaa kolledži Telemaatika ja arukate süsteemide eriala esmakursuslaste väljalangemise ennustamiseks, mis suudaksid juba sisseastumise etapil tulevasi tudengeid iseloomustada ning lubaksid sisseastumisvestluse tulemusi korrigeerida. Töö käigus uuriti struktureeritud ja struktureerimata andmed: kandidaadi avalikud üld- ja õppeandmed, sisseastumisvestluse, essee ja sisseastumisküsitluse andmed, mida võib koguda sisseastumiseprotsessis iga õppeasutus. Autorile teadaolevalt selline kompleksne struktureeritud ja struktureerimata andmetel põhinev uuring on unikaalne ning erineb samal teemal tehtud töödest. Statistiliste meetoditega andmete analüüsimisel leiti tudengite väljalangemist mõjutavad faktorid. Töö teoreetilises osas anti käsitletud probleemi teemal Eestis läbi viidud varasematest uurimustest ülevaade ja töö käigus olid kinnitatud mõned varem leitud väited. Töö praktilises osas leiti mudelite ennustustäpsused. Lõpuks pakuti soovitused sisseastumisprotseduuri efektiivsuse tõstmiseks.

Töö tulemuste põhjal saab teha järeldust, et saadud mudelite ristvalideeritud ennustusvõime hinnangud ületavad Eestis varem avaldatud enne õpingute alustamist kogutud andmetel põhinevaid uurimistulemusi. Leitud mudelite mõjukad tunnused on Keskhariduse lõputunnistuse keskmine hinne, Linna suurus („väike“), sisseastuja elukoht (Ida-Virumaa), vanus, essee leksikaalne rikkus ja vestluse punktid, kusjuures viimasel on väga nõrk mõju. Parimad mudelid on tugivektor-masin (F1=0.7825) ja AdaBoost (F1 = 0.7671).

Küsitlus on kõige parem viis informatsiooni saamise ja ümbertöötlemise jaoks. Oskuslikult valitud küsimused võimaldavad aidata väljalangemise klassifitseerimisel. Töö käigus oli leitud, et küsimused eriala tundmisest, eeloskuste enesehindamisest, sotsiaaloskustest ja isiklike andmete kombinatsioon on ennustusmudeli tähtsad tunnused. Nendel tunnustel parimad mudelid on logistiline regressioon (F1=0.7805), PLS-DA (F1=0.7464), tugivektor-masin (F1=0.7354). Küsitlusel põhinev andmestik

nõuab edaspidist arendamist liiga väikese objektide arvu ja palju oma vahel seotud küsimuste tõttu.

Töö käigus oli tõestatud, et töös analüüsitud tudengite esseede arvulised tunnused täiendavad ja rikastavad mudeleid. Teksti arvulised tegurid on tähtsamad kui küsitluse või üldandmed. Vestluse andmed on kõige mitteolulisemad väljalangemise ennustamisel.

TalTech Virumaa kolledži sisseastumisprotseduuri efektiivsuse tõstmiseks autor pakub kõigepealt lisada motivatsioonikirja, sest esseede tekstide baasil tuletatud arvulised tunnused suurendasid mudelite ennustusvõimet. Teiseks, kasutada üld- ja haridusandmeid, motivatsioonikirja ja küsitlusele vastuseid individuaalsete vestluse küsimuste koostamiseks ja vestluse hindamisel arvestada antud küsimusetele kandidaadi vastuseid. Kolmandaks, tõsta vestluse lävendi kuni 6 punktini ehk 60% kogusummast. Neljandaks, kohustuslikult fikseerida kandidaadi matemaatika eksami tulemus, sest antud väärtus ei saa asendada mingi teisega. Viiendaks, keskhariduse tunnistuse keskmise hinde tõstmine 3.8-st kuni 3.9 ei avalda mõju väljalangemise probleemile. Kolm viimast pakkumist kehtivad ainult konkreetse õppeasutuse puhul ja neid ei saa üldistada.

Kasutatud kirjandus

[1] „Eesti tööturg täna ja homme“, SA Kutsekoda, 2016. [Võrgumaterjal] https://oska.kutsekoda.ee/wp-content/uploads/2017/02/Eesti_tooturg.pdf [Kasutatud 12 2021].

[2] N. Maksimova, A. Pentel, O. Dunajeva, “Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study”, Springer, 719−726, 2020. DOI: 10.1007/978-3-030-68201-9_70.

[3] O. Must, A. Must, “Kõrgkoolist väljalangevus ja üliõpilaste enesemäärtlus”, Tartu Ülikool, 2017.

[4] V. Tinto, “Dropout from Higher Education: A theoretical Synthesis of Recent Research”, Review of Educationsl Research, Vol.45, No. 1, pp. 89-125, 1975.

[5] K. Kori, M. Pedaste, E. Tõnisson, T. Palts, H. Altin, R. Rantsus, R. Sell, K. Murtazin, T. Rüütmann, “First-year dropout in ICT studies”, 2015 IEEE Global Engineering Education Conference (EDUCON), IEEE Digital Library, 2015.

[6] O. Liblik, “Klassifitseerimise algoritmi abil e-õppe süsteemis tudengite väljalangemise ennetamine informaatika aine näitel”, magistritöö, Tallinna Tehnika Ülikool, 2016.

[7] U. Brenda, “TTÜ tudengite väljalangemise ennustamine: tõenäosuse arvutamine masinõppemeetodite abil ning tulemuste kuvamine veebirakenduses”, bakalaureuse töö, Tallinna Tehnika Ülikool, 2017.

[8] T. Ketevan, “Predicting Academic Performance From Admission Scores and Application Data –A Case Study”, Master’s Thesis, University of Tartu, 2019.

[9] O. Dunajeva, A. Pentel, N.Maksimova, “COVID-19’s Impact on the Quality of Educational Process and the Academic Performance as Viewed by IT Students ,A Case Study in Text Mining”, Springer, ICL 2021, Volume 1, LNNS 389. DOI:10.1007/978-3-030-93904-5_42. (ilmumisel)

[10] Prem S. Mann, “Introductory Statistics”, Seventh Edition, New York, Wiley, 2010.

[11] V. Egoshin, S. Ivanov, N. Savvina, G. Kapanova, L. Zhamaliyeva, A. Grjibovski, “Analysis of Categorical Variables Using R. Ekologiya cheloveka”, Human Ecology, 1, pp. 51-64, 2019.

[12] V. Bewick, L. Cheek, J. Ball, “Statistics review 8: qualitative data - tests of association”, Crit Care, 8:46–53, 2004.

[13] Julien I.E. Hoffman, “Exploratory Descriptive Analysis”, Basic Biostatistics for Medical and Biomedical Practitioners, pp. 49-89, 2019.

[14] R. Pearson, “The GoodmanKruskal package: Measuring association between categorical variables”, 18/03/2020. [Võrgumaterjal] https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html. [Kasutatud 12 2021]

[15] J.H. McDonald, “Handbook of Biological Statistics”, (3rd ed.), Sparky House Publishing, Baltimore, Maryland, pp 157-164, 2014.

[16] L. A. Goodman, W. H. Kruskal, “Measures of association for cross classifications”, Journal of the American Statistical Association, 49(268), 732–764, 1954. doi:10.2307/2281536

[17] “Goodman and Kruskal's gamma using SPSS Statistics”, Laerd Statistic, [Veebimaterjal] https://statistics.laerd.com/spss-tutorials/goodman-and-kruskals-gamma-using-spss-statistics.php [Kasutatud 12 2021]

[18] H. Y. Kim, Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test, Restorative Dentistry & Endodontics, 42(2): 152–155, 2017. 10.5395/rde.2017.42.2.152.

[19] A. Aasa, K.Lõhmuse, M.Espenberg, “Mitteparameetriliste meetodite praktikumi juhend”, Tartu Ülikool, 2020 [Veebimaterjal] http://aasa.ut.ee/statistika/prax_06.html [Kasutatud 12 2021]

[20] T. Kaart, “Binaarsete tunnuste analüüsimeetodid”, EMÜ VLI, 2012. [Veebimaterjal] http://ph.emu.ee/~ktanel/bin_tunnuste_analyys/bin_tunnuste_analyys.pdf [Kasutatud 10 2021].

[21] M. Kuhn, K. Johnson, “Applied Predictive Modeling”, Springer, New York, 2013.

[22] L. Simon, D. Young, I. Pardoe, “Applied Regression Analysis”, The Pennsylvania State University, 2018. [Võrgumaterjal] https://online.stat.psu.edu/stat462/node/96/ [Kasutatud 12 2021].

[23] Williams, G. Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery. New York : Springer, 2011.

[24] Y. Huang, L. Li, “Naive Bayes classification algorithm based on small sample set", 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 34-39, 2011. doi: 10.1109/CCIS.2011.6045027.

[25] M. Kuhn, “The Caret package”, 27.03.2019. [Võrgumaterjal] https://topepo.github.io/caret/model-training-and-tuning.html [Kasutatud 12 2021].

[26] F.Maleki, N. Muthukrishnan, K.Ovens, C. Reinhold, R. Forghani,” Machine Learning Algorithm Validation”, Neuroimaging Clinics of North America, 30(4):433-445, 2020. DOI: 10.1016/j.nic.2020.08.004.

[27] Y. Sasaki, “The truth of the F-measure”, 2007.

[28] K. Remm, J. Remm, A. Kaasik, „Ruumiliste loodusandmete statistiline analüüs“. E-õpik, TÜ Ökoloogia ja Maateaduste Instituut, Tartu, 2012. http://hdl.handle.net/10062/26456.

[29] J. Cohen," A coefficient of agreement for nominal scales", Educational and Psychological Measurement, 20: 37–46, 1960.

https://www.sciencedirect.com/science/article/pii/B0123693985000955

[30] K. Malterud, V. Siersma, A. Guassora, " Sample size in qualitative interview studies: guided by information power", Qual Health Res. 2015;26:1753–1760. doi: 10.1177/1049732315617444.

[31] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A.B. Dris, N. Alzakari, A. Abou Elwafa, H. Kurdi, "Impact of Dataset Size on Classification Performance", An Empirical Evaluation in the Medical Domain. Appl. Sci. 2021, 11, 796. https://doi.org/10.3390/app11020796

[32] Zhang, Y., Ling, C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater 4, 25 (2018). https://doi.org/10.1038/s41524-018-0081-z

[33] J. Brownlee, "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset", Machine Learning Mastery, 15.08.2020. [Võrgumaterjal] https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ [Kasutatud 10 2021].

[34] H. He and E. A. Garcia, Learning from Imbalanced Data, in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009. DOI: 10.1109/TKDE.2008.239.

[35] Li, K.; Zhou, G.; Zhai, J.; Li, F.; Shao, M. Improved PSO_AdaBoost Ensemble Algorithm for Imbalanced Data. Sensors 2019, 19, 1476. https://doi.org/10.3390/s19061476.

[36] Haridus ministeerium, Analüüs ja ettepanekud eesti keele õppe tõhustamiseks põhikoolis, lk.3, 2014. [Võrgumaterjal] https://www.hm.ee/sites/default/files/analuus_ja_ettepanekud_eesti_keele_oppe_tohustamiseks_pohikoolis.pdf [Kasutatud 11 2021].

[37] A. Vermeer," Coming to grips with lexical richness in spontaneous speech data", Language Testing, 17, 65-83, 2020. Doi: 10.1177/026553220001700103.

[38] McCarthy P. M. Vocd: "A theoretical and empirical evaluation", Language Testing, 24(4): 459-488, 2007.

[39] R. Kasik, "Sissejuhatus tekstiõpetusse", Tartu, 2007.

[40] J.P. Kincaid, R.P. Jr. Fishburne, R.L. Rogers, B.S. Chissom (February 1975), "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel", Institute for Simulation and Training, 56, 1975. https://stars.library.ucf.edu/istlibrary/56.

[41] D. Spadaro, A. Robinson, D. Lawrence , L. Smith, " Assessing Readability of Patient Information Materials", American journal of hospital pharmacy, 37(2): 215-221, 1980. DOI: 10.1093/ajhp/37.2.215.

[42] R. Rambo, "The reading level of your writing", English Composition 1,2019. [Võrgumaterjal] http://facultyweb.ivcc.edu/rrambo/eng1001/reading_level.htm [Kasutatud 10 2021].

Summary

The purpose of this graduation thesis was to use machine learning methods to create prediction models for TalTech Virumaa college's Telematics and Smart Systems programme's freshmen dropout rates, which could help to characterise prospective students even at the admission stage and let the admission committee correct the admission interview's results. Throughout the paper the following structured and unstructured data were researched: the candidate's public general and study data; admission interview, essay, and admission survey data, which can be collected at any institution during the admission process. As far as the author knows, such complex research based on structured and unstructured data is unique and differs from other works on this topic. The factors that influence students' dropping-out were found by analysing data with statistical methods. The theoretical part of the thesis consists of an overview of the earlier, carried out in Estonia research regarding the handled problem, as well as the confirmation of some earlier found claims. The practical part of the thesis consists of finding the models' prediction accuracies. The final part of the paper offers recommendations for increasing admission process's efficiency.

Based on the research results it can be concluded that the evaluations of models' cross-validated prediction ability exceed earlier published in Estonia research results based on collected data from before starting studies. The found models' important attributes are as follows: high school diploma's GPA, town's size (small), prospective student's place of residence (Ida-Virumaa), age, essay's lexical diversity and admission interview's points, where the latter has a very weak influence. The best models are support-vector machine (F1=0.7825) and AdaBoost (F1 = 0.7671).

The survey is the best way to get and process information. Capably chosen questions allow to classify dropouts. Throughout the paper it was found that the most influential questions are about field knowledge and self-assessment of one's abilities, including social abilities. The best models of that data are Logistic Regression (F1=0.7805), PLS-DA (F1=0.7464) and support-vector machine (F1= 0.7354).

During the research, it was proven that the quantifiable characteristics of students' essays analysed in the paper complete and enrich the models. Text's quantifiable factors are more important than those of the survey or general data. The interview's data are the most non-important for predicting dropping-out.

The dataset based on the survey requires a further development because of a too small number of the objects and many connected to each other questions.

To increase TalTech Virumaa college's admission process's efficiency the author proposes, first, adding a motivation letter because the quantifiable characteristics derived from the essays' texts increased models' prediction ability. Second, using general and study data, motivation letter and answers to the survey to create questions for the individual interview and taking into consideration the candidate's answers given to the questions during the interview's grading. Third, raising the interview's threshold level to 6 points, or 60% of the whole sum. Fourth, compulsorily fixate the candidate's math exam result because the given value can't be replaced by anything else. Fifth, high school diploma's GPA's increase from 3.8 to 3.9 doesn't reveal the influence on the dropout problem. The last three propositions only apply to particular institutions and can't be generalized.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[8]

Mina, Natalja Maksimova

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „IKT eriala üliõpilaste varajase väljalangemise ennustamine masinõppe meetodite abil“, mille juhendaja on Olga Dunajeva
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

01.01.2022.

8 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

Lisa 2: 2019. ja 2020. aastate SAIS'i ja vestluse andmete tehtud teisendused ja kirjeldus (andmestik B)

SAIS süsteemi andmetest oli saadud edaspidiseks analüüsiks 15 tunnust:

Õppevorm, Sugu, Kodakondsus, Sünniaeg, E-post, Aadress: rahvastikuregistrist, Telefon, Õppeasutuse nimi, EHISe õppetase, Õppekeel, Tunnustus, Lõputunnistuse keskmise hinne üle kõigi keskkooli ainete, Konkursipunktid vestluse eest, Vestluse kuupäev.

Nendest andmetest oli loodud andmestik B. Pärast teisendust kirjeldab andmekogumit 16 tunnust, demograafia, kooli ja vestluse tunnused, kus sulgudes on määratud esialgne tunnus ja valemi seletus: *Aasta, Vorm, Sugu, Kodakondsus, Vanus* (⇐ Sünniaeg), *Kuu* (⇐ Sünniaeg), *EmailHost* (⇐ E-post, kas com, ru või ee), *ametlikEmail* (⇐ E-post, kas nimi.perekonnanimi@host), *Linna_suurus* (⇐ Aadress rahvastikuregistrist + Eesti linnade ja valdade elanike arv[9] = järjestatud tunnus klassidega L (elanike arv >30000), M (30000-5001), S (<5000)), *Maakond* (⇐ Aadress: rahvastikuregistrist), *Koolikeel* (⇐ Õppekeel), *Kutsekool* (⇐ EHIS õppetase), *Kooli_tase* (⇐ Õppeasutuse nimi + Kooli matemaatika riigieksami pingerida = järjestatud tunnus I - IV (pingerida oli jaotud kvartiilideks ja kui koolinimetus pingireas puudus, siis määramata)), *Kkh* (=Lõputunnistuse keskmine hinne üle kõigi keskkooli ainete), *VeP* (=Konkursi punktid vestluse eest), *V_suvel* (⇐Vestluse kuupäev, kas suvel või muul ajal).

```
'data.frame':   89 obs. of  17 variables:
 $ Aasta        : int  2020 2020 2020 2020 2020 2020 2020 2020
 $ Vorm         : Factor w/ 2 levels "kaug","päev": 2 1 1 1 1
 $ Kuu          : int  8 5 10 7 3 6 8 11 3 10 ...
 $ Vanus        : int  19 20 49 37 33 19 19 20 19 33 ...
 $ emailHost    : Factor w/ 3 levels "com","ee","ru": 1 3 2 1
 $ ametlikEmail : Factor w/ 3 levels "ei","jah","segane": 2 3
 $ Sugu         : Factor w/ 2 levels "m","n": 1 1 1 1 1 1 1 1
 $ Pass_Est     : int  1 1 1 1 0 1 1 1 0 1 ...
 $ Linna_suurus : Factor w/ 3 levels "L","M","S": 1 1 2 1 1 2
 $ Ida.Viru     : int  1 1 1 0 1 1 1 1 1 0 ...
 $ Kool_keel_est: int  0 0 0 1 0 0 1 0 0 1 ...
 $ Kutsekool    : int  1 1 1 1 1 1 0 1 1 0 ...
 $ Kooli_tase   : Factor w/ 5 levels "I","II","III",..: 5 5 5
 $ KKp          : num  4.24 4.67 4.27 3.74 4.27 4.27 4.11 4.19
 $ VeP          : num  8 5.5 10 5.38 7.25 ...
 $ Vest_suvel   : int  1 1 1 1 1 1 1 1 1 0 ...
 $ y            : int  0 0 0 0 0 0 0 0 1 0 ...
```

[9] Eesti linnade ja valdade liit, https://www.elvl.ee/elanike-arv

Andmestiku kirjeldav statistika

```
      Kuu             Vanus            KKp             VeP
 Min.   : 1.00   Min.   :19.00   Min.   :3.600   Min.   : 5.000
 1st Qu.: 4.00   1st Qu.:20.00   1st Qu.:3.880   1st Qu.: 6.100
 Median : 7.00   Median :24.00   Median :4.110   Median : 7.250
 Mean   : 6.82   Mean   :27.06   Mean   :4.145   Mean   : 7.171
 3rd Qu.:10.00   3rd Qu.:34.00   3rd Qu.:4.380   3rd Qu.: 8.300
 Max.   :12.00   Max.   :49.00   Max.   :5.000   Max.   :10.000
```

2019. ja 2020. aastate sisseastujate keskmine vanus on 27 aastat, keskhariduse lõputunnistuse keskmine hinne on 4.145 ja vestluse keskmine punktide arv on 7.2.

```
 emailHost ametlikEmail Sugu   Linna_suurus Kooli_tase  Aasta    Pass_Est Ida.Viru Kool_keel_est
 com:65    ei     :22    m:73   L:30         I  : 4     2019:45   0:11     0:22     0:45
 ee : 9    jah    :50    n:16   M:40         II :10     2020:44   1:78     1:67     1:44
 ru :15    segane:17            S:19         III:14
                                             IV : 8
                                             MT :53

 Kutsekool Vest_suvel
 0:42      0:30
 1:47      1:59
```

Andmekogumi B 89 sisseastutajatest on 82% mehed ja 18% naised. Sessioonõpe vorm on populaarsem, sessioonõpe vormi osakaal on 64%. Enamik sisseastutajatest 67 (75%) on pärit Ida-Virumaalt, samas maakonnas asub TalTech Virumaa kolledž. Umbes 53% sisseastujatest lõpetas kutsekooli.

Andmestik B on esitatud ka dihhotoomsete tunnuste kaudu, millel on vaid kaks võimalikku väärtust. Dihhotoomsel kujul andmestik sisaldab 43 tunnust:

```
'data.frame':   89 obs. of  43 variables:
 $ Pass_Est          : int  1 1 1 1 0 1 1 1 0 1 ...
 $ Ida.Viru          : int  1 1 1 0 1 1 1 1 1 0 ...
 $ Kool_keel_est     : int  0 0 0 1 0 0 1 0 0 1 ...
 $ Kutsekool         : int  1 1 1 1 1 1 0 1 1 0 ...
 $ Vest_suvel        : int  1 1 1 1 1 1 1 1 1 0 ...
 $ Aasta_2019        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Aasta_2020        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Vorm_kaug         : int  0 1 1 1 1 1 0 0 1 1 ...
 $ Vorm_päev         : int  1 0 0 0 0 0 1 1 0 0 ...
 $ emailHost_com     : int  1 0 0 1 0 1 1 1 1 1 ...
 $ emailHost_ee      : int  0 0 1 0 0 0 0 0 0 0 ...
 $ emailHost_ru      : int  0 1 0 0 1 0 0 0 0 0 ...
 $ ametlikEmail_ei   : int  0 0 0 1 0 0 0 1 0 0 ...
 $ ametlikEmail_jah  : int  1 0 1 0 0 1 1 0 0 1 ...
 $ ametlikEmail_segane: int  0 1 0 0 1 0 0 0 1 0 ...
 $ Sugu_m            : int  1 1 1 1 1 1 1 1 0 1 ...
 $ Sugu_n            : int  0 0 0 0 0 0 0 0 1 0 ...
 $ Linna_suurus_L    : int  1 1 0 1 1 0 0 1 1 1 ...
 $ Linna_suurus_M    : int  0 0 1 0 0 1 1 0 0 0 ...
 $ Linna_suurus_S    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Kooli_tase_I      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Kooli_tase_II     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Kooli_tase_III    : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Kooli_tase_IV     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Kooli_tase_MT     : int  1 1 1 1 1 1 0 1 1 1 ...
 $ Kuu_Ikv           : int  0 0 0 0 1 0 0 0 1 0 ...
 $ Kuu_IIkv          : int  0 1 0 0 0 1 0 0 0 0 ...
 $ Kuu_IIIkv         : int  1 0 0 1 0 0 1 0 0 0 ...
 $ Kuu_VIkv          : int  0 0 1 0 0 0 0 1 0 1 ...
 $ Vanus_18_20       : int  1 0 0 0 0 1 1 0 1 0 ...
 $ Vanus_20_25       : int  0 1 0 0 0 0 0 1 0 0 ...
 $ Vanus_25_30       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Vanus_30_35       : int  0 0 0 0 1 0 0 0 0 1 ...
 $ Vanus_35_50       : int  0 0 1 1 0 0 0 0 0 0 ...
 $ KKp_3.5_3.7       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ KKp_3.7_4.1       : int  0 0 0 0 1 0 0 0 0 0 ...
 $ KKp_4.1_4.5       : int  1 0 1 0 1 1 1 1 1 0 ...
 $ KKp_4.5_5         : int  0 1 0 0 0 0 0 0 0 0 ...
 $ VeP_v6            : int  0 1 0 1 0 0 0 0 1 0 ...
 $ VeP_6_7           : int  0 0 0 0 0 1 1 0 0 0 ...
 $ VeP_7_9           : int  1 0 0 0 1 0 0 1 0 1 ...
 $ VeP_9_10          : int  0 0 1 0 0 0 0 0 0 0 ...
 $ y                 : int  0 0 0 0 0 0 0 0 1 0 ...
```

Lisa 3: Vestluse käik ja teemad

Telemaatika ja arukad süsteemid

Vestlusel on võimalik käia ainult üks kord. Vestluse käigus esitatakse küsimusi ja tutvutakse üliõpilaskandidaadi motivatsiooni ning eelnevate teadmistega eriala kohta.

Vestluse käigus hinnatavad valdkonnad/teemad:

1. Huvi eriala vastu, õppekava valiku põhjendus Kandidaat põhjendab selgelt oma huvi Telemaatika ja arukate süsteemide õppekaval õppimise vastu, tal on olemas ülevaade õppekava sisust ja peaerialadest; põhjendab peaeriala eelistust ja seob selle oma arengueesmärkidega.
2. Tulevikuväljavaated, valmidus erialal töötamiseks Kandidaadil on selge arusaam tulevastest tööalastest võimalustest, ta toob näiteid spetsialiseerumisest pikemas perspektiivis, võimalikest tegevusaladest ja ametikohtadest; ettekujutus tööalasest karjäärist ja kuidas Telemaatika ja arukate süsteemide õppekava läbimine sellele kaasa aitaks; vastab küsimusele: "Kellena soovid töötada 10 aasta pärast?".
3. Õpivalmidus, suutlikkus õpingute lõpetamiseks Kandidaadil on valmisolek omandada infotehnoloogia ja automatiseerimise alased oskused ja teadmised; ootused õppe osas ehk arusaam omandatavatest teadmistest ja oskustest; ta oskab hinnata oma ajalisi ja rahalisi võimalusi täiskoormusega õppimiseks; on läbi mõelnud õppekava läbimist takistavad võimalikud töö- ja/või eraelulised riskid ning nende maandamise võimalused.
4. Eelnev haridus ja töökogemus Kandidaadil on eelnev haridus või töökogemus ning oodatav lisandväärtus Telemaatika ja arukate süsteemide õppekava läbimisest. Kohapeal võidakse paluda lahendada suuliseid matemaatika ülesandeid.

Vestluse hindamiskriteeriumid

Vestluse iga komponendi eest on võimalik saada kuni 2,5 punkti. Vestluse positiivne skaala 5–10 punkti, negatiivne skaala 0–5 punkti. 0 punkti saab kandidaat, kes ei ilmu vestlusele.

Lisa 4: Esseede andmete kirjeldus ja tehtud teisendused

Esimesel etapil oli tõlgitud esseed inglise keelde. Antud etapp oli teostatud *Pythoni* olemasoleva tasuta paketi *DeepL*[10] abil, mis pakub palju tööriistu erinevate keelte vaheliste tõlgete tegemiseks, kasutades mitmeid masintõlkijaid nagu google, yandex, microsoft jt, milledest autoril oli valitud Google Translate. Tõlkimise näide on esitatud Joonisel 33.

```
Я надеюсь начать подрабатывать строго по моей специальности еще до окончания самой
учёбы. Например – помогать людям переустанавливать и настраивать операционную систему,
может даже настроить сервер или локальную сеть, создавать сайты, обслуживать уже
имеющиеся сайты (дорабатывать, расширять, сделать отдельный сайт для дочерней фирмы,
сделать веб-приложение для сайта).

I hope to start earning extra money strictly in my specialty even before I finish my
studies. For example, helping people reinstall and configure the operating system, can
even set up a server or local network, create sites, maintain existing sites (modify,
expand, make a separate site for a subsidiary, make a web application for the site).
```

Joonis 33. Venekeelse essee osa tõlkemise näide.

Tekstide eeltöötluse protsess koosnes neljast osast: 1) puhastamine (*normalize*), 2) sõnade eraldamine (*tokenize*) ja kuna teksti masinõppes kasutamiseks on vaja vähendada teksti andmete variatiivsust, siis järgmised sammud aitavad seda teha: 3) lemmatiseerimine (*lemmatize*), 4) stoppsõnade eraldamine (*romove Stop Words*). Eeltöötluse protsessi tulemuse näidis on esitatud Joonisel 34.

```
[1 lause]: I hope to start earning extra money strictly in my specialty even before I
finish my studies.

[korpuse osa]:"rus f session yes.1 :

I hope to start earning extra money strictly in my specialty even before I finish my
studies"

[tokanize]: "I" "hope" "to" "start" "earning" "extra" "money" "strictly" "in" "my"
"specialty" "even" "before" "I" "finish" "my" "studies" [17 sõna]

[lemmatize]: "earning" -> "earn"; "studies" -> "study"

[stop words remove]: "hope" "start" "earn" "extra" "money" "strictly" "specialty"
"even" "finish" "study" [10 sõna]
```

Joonis 34. Essee teksti eeltöötluse protsessi näide.

Tekstidest moodustatud korpus sisaldab 67 dokumenti ja pärast numbrite, stoppsõnade eemaldamist ja lemmatiseerimist oli esseede sõnade koguarv 9968, unikaalsete sõnade arv 1718. Iga essee sõnade koguarv on vahemikus 20 kuni 314 keskmisega 141 ja

[10] https://deep-translator.readthedocs.io/en/latest/

standardhälbega 96. Kordumatute sõnade arv igas essees oli 18 kuni 183 keskmisega 104 ja standardhälbega 33.

Sõnade arvud, unikaalsed sõnad ja sagedasemad sõnad leiti termin-dokument maatriksite abil, olulisemad sõnad määrati tf-idf (termini sagedus - dokumendi pöördsagedus) tehnikaga. Kõige sagedasemad ja tf-idf-i olulisemad sõnad üliõpilaste esseedes teemal "Minu nägemus tulevasest erialasest tööst" on kujutatud Joonisel 35 sõnapilvedena.

(a) (b)

Joonis 35. (a) kõige sagedamate ja (b) tf-idf olulisemate sõnapilvid.

Edaspidi ennustusmudelites kasutamiseks oli loodud järgmised arvulised tunnused:

Keele leksikaalne rikkus (ingl *lexical diversity*) - arvutab leksikaalset mitmekesisust, tuginedes unikaalsete sõnade arvule ja dokumendi pikkusele. See on kasulik esinejate või kirjanike keeleoskuse või dokumentides väljendatud ideede keerukuse analüüsimiseks. Teksti leksikaalse mitmekesisuse arvutamiseks kasutatakse erinevaid mõõdikuid, selles töös oli valitud Uberi indeks[11], mis arvutatakse valemi 3 järgi:

$$U = \frac{(logN)^2}{logN - logV}, \text{ kus} \quad (3)$$

N – sõnade koguarv (ingl *tokens*), *V* – unikaalsete sõnade arv (ingl *types*). Mida suurem on U, seda rikkamaks võib pidada teksti autori sõnavara.

Uberi indeks oli valitud, sest see on suhteliselt sõltumatu teksti pikkusest [37] ja mõõdetakse sarnases skaalas koos teiste loodud teksti iseloomustavate arvuliste tunnustega. Uberi index on Maasi indeksi pöördväärtus, viimane ei sõltu ka kirjaliketest žanritest [38].

Teksti mõistetavus ehk loetavus (ingl *readability*) – kuivõrd tekst on keeruline lugejale. See on kasulik ega õppeteksti mõistetavuse taseme määramiseks, kuid esseede

[11] Calculate lexical diversity, Quanteda, http://quanteda.io/reference/textstat_lexdiv.html

uuringutes. Kuna hea essee tunnus on „sõnastus on selge ja ladus ning seda on kerge lugeda“[12] , siis on kasulik ka arvutada teksti mõistetavust, mis seostub eeskätt teema selgekstegemisega, sobivate näidete valikuga, sõnavara konkreetsusega, aga sõltub ka lugeja teadmistest ja arusaamisvõimest [39]. Siin kasutati Flesch-Kincaid Grade Level [40] arvutusvalemit (4):

$$0.39\,\frac{\text{unikaalsete sõnade arv tekstis}}{\text{kokku lausete arv}} + 11.8\,\frac{\text{silpide arv}}{\text{unikaalsete sõnade arv tekstis}} - 15.59 \qquad (4)$$

Valemi tulemusena on USA haridussüsteemi järgi teksti mõistmiseks vajalik haridustase (vt Tabel 12).

Tabel 12. Teksti loetavuse taseme vastavus USA haridustasemele [41].

Flesch-Kincaid Grade Level	**Haridustase**	**Selgitus**	**Keskm. sõnade arv lauses**	**Silpide arv 100 sõnades**
[5; 6)	US 5.klass	väga lihtne	8	123
[6; 7)	US 6. klass	lihtne	11	131
[7; 8)	US 7. klass	üpris lihtne	14	139
[8;10)	US 8.-9 klassid	standardne	17	147
[10; 12)	US 10.-12. klassid, gümnaasium	üpris keeruline	21	155
[13; 16)	US kolledž	keeruline	25	167
[16; …)	US kolledži lõpetaja	väga keeruline	29	192

Nt essee teksti mõistetavus 7.2 vastab 7-nda klaasi tasemele. 7-nda klassi tasemele teksti mõistetavus tähendab tihti suure hulka lühi sõnade ja lausete kasutamist [42]. Antud hinnangu valemi esimene liidetav sõltub sõnade ja lausete pikkusest, selle tõttu lühilausete ühendus või põimlausete osemasolu suurendab essee mõistetavuse tase.

Järgmised karakteristikud arvutati originaalkeeles esseedel.

Teksti vesi – „vesise” näitaja (Valem 5), stoppsõnade protsent kontekstis.

$$vesi = \frac{stoppsõnade\ arv}{unikaalsete\ sõnade\ arv} \cdot 100\% \qquad (5)$$

[12] Kirjalike üliõpilastööde struktuur ja vormistamine Tallinna Tervishoiu Kõrgkoolis,lk.4

Teksti „rämps“ – kõige sagedamini esinevate sõnade arvu ja teksti sõnade koguarvu suhe (Valem 6), mis peegeldab otsingu märksõnade (vt Joonis 36) arvu tekstis. Mida rohkem märksõnu tekstis, seda suurem on selle „rämpsus“. Märksõnadeks loetakse selles töös esseede kõige sagedamaid ja essee pealkirja sisaldavaid sõnu (erinevates vormides).

```
keyWords <- c("it","telemaatika",
              "tulevik","tuleviku","tulevikus",
              "aruka","iot","eriala", "süsteemide",
              "töö","kindlasti", "arukad","arukat",
              "süsteemi","süsteem","süsteemid","ит",
"телематика","будущий","будущее","будущие","умные",
"умный","умное","умная","системы","система","иот",
"работа","работу","работы","будущем","очень")
```

Joonis 36. Essee märksõnade loetelu rämpsuse valemi jaoks.

$$rämpsus = \frac{märksõnade\ koguarv}{ilma\ stoppsõnadeta\ unikaalsete\ sõnade\ arv} \cdot 100\% \qquad (6)$$

Plagiaat – võõra teose või selle osade avaldamine ilma autorile viitamiseta. Iga essee plagiaadi protsent oli kontrollitud Moodle süsteemisse lisatud Urkund laienduse kaudu.

Lisa 5: Küsitluse küsimused

Üldandmed

1. Ees- ja perekonnanimi. Nominaalne tunnus. Tekstkast.
2. Sünniaeg. Diskreetne tunnus. Valik kalendrist.
3. Elukoht. Nominaalne tunnus. Üks valik 17 väärtusest: {Eesti maakonnad; Muu}.
4. Kui palju aega keskmiselt kulub Teil kolledži jõudmiseks? Nominaalne tunnus. Üks valik 5 klassist: {alla 15 min; 15-30 min; 30 min - 1 tundi; 1 - 2 tundi; üle 2 tundi}.
5. Hinnake oma keelteoskust (1 – ei oska üldse, 5 – oskan suurepäraselt) Palutakse hinnata 5.1 eesti, 5.2 vene, 5.3 inglise keeled.

Eriala

6. Hinnake oma informeeritust Telemaatika ja arukate süsteemide eriala kohta (1 - ei ole üldse teadlik sellest erialast, 5 - väga teadlik sellest erialast).
7. Millega on telemaatika seotud? Hinnake iga punkti (1 – üldse ei ole seotud, 5 – on väga seotud). Pakutakse hinnata 8 punkti (Televisioon, Meedia, Andmete analüüs, Programmeerimine, Nutimaja, Masinõpe, Andmeedastus, Andur).
8. Millised õppeained on eriala Telemaatika ja arukad süsteemid õppekavas? Mitmene valik. Pakutaks valida mis tahes punkti 7 variandist {Telekommunikatsiooni alused, Raadiotehnika / elektroonika alused, Andmeanalüüs, Programmeerimine, Arukad süsteemid, Andmeteadus ja masinõpe, Digitaalne meedia}.
9. Millise valdkonna õppeained Teid kõige rohkem huvitavad? Mitmene valik {Tootmise automatiseeritud süsteemid, Arukad süsteemid, Programmeerimine, 3D modelleerimine, Andmeteadus, Side ja küberturvalisus}.
10. Millise õppevormi valite? Binaarne tunnus: {päevaõpe; sessioonõpe}.
11. Kas Teil on lõpetatud või lõpetamata haridus erialal Telemaatika ja arukad süsteemid või sellele lähedasel erialal? Nominaalne tunnus. Üks valik: {Puudub; Lõpetamata kutse- või keskeriharidus; Lõpetamata kõrgharidus või

rakenduskõrgharidus; Lõpetatud kutse- või keskeriharidus; Lõpetatud kõrgharidus või rakenduskõrgharidus}.

12. Täpsustage lõpetatud / lõpetamata hariduse eriala. Nominaalne tunnus. Unikaalne tekst.
13. Kas Telemaatika ja arukate süsteemide eriala on teie esimene valik edasiõppimiseks? Binaarne tunnus: {jah; ei}.
14. Kas teil on töökogemusi telemaatika erialal või sellele lähedasel erialal? Binaarne tunnus: {jah; ei}.
15. Kui vastasite „jah“, siis täpsustage, mitu aastat. Nominaaltunnus. Unikaalne tekst. Kui pole täiendatus, siis 0.
16. Hinnake oma eelteadmisi, mis on vajalikud edasiseks õppimiseks (1 - vajalikud eelteadmised puuduvad täielikult, 5 - suurepärased eelteadmised) 16.1 matemaatika, 16.2 füüsika, 16.3 programmeerimine, 16.4 inglise keel, 16.5 praktiline kogemus IT valdkonnas, 16.6 automaatika, 16.7 praktiline kogemus automatiseerimise valdkonnas.
17. Hinnake oma valmisolekut ainete iseseisvaks õppimiseks (1 - pole üldse valmis, 5 - täiesti valmis).
18. Mitu tundi päevas olete valmis õppimisele pühendama? Järjestatud tunnus. Üks valik {0 – 2; 2-4: 4-6: >6}.
19. Kui palju Teie arvates aitab kolledžidiplom soovitud töökoha saamisel? (1 - ei aita üldse, 5 - aitab palju).
20. Kas Teie pere / lähedased toetavad Teie otsust jätkata õpinguid kolledžis? (1 - ei toeta üldse, 5 - toetab väga).
21. Kui suur on Teie ootus, et kolledžis õppimine annab palju juurde Teie arengule valitud õppevaldkonnas? (1- ei oota, et annab palju juurde, 5 - ootan, et annab väga palju juurde) .

Isiklikud omadused

22. Hinnake oma kallakut järgmiste teaduste poole (1 - üldse mitte, 5 - väga tugev) 22.1 Humanitaarteadused 22.2 Reaalteadused
23. Kas teile meeldib töötada meeskonnas? (1 – ei meeldi üldse, 5 – väga meeldib)
24. Millist rolli eelistaksite meeskonnatöös? Nominaaltunnus. Üks valik: {liider; täitja; vaatleja}.

25. Mida te eelistaksite raskuste tekkimisel? Nominaaltunnus. Üks valik: {Iseseisvalt hakkama saada; Küsida abi sõprade, sugulaste, tuttavate käest; Pöörduda spetsialistide poole; Ootan, et olukord ise laheneb}.
26. Tavaliselt Te kaldute ennast? Nominaaltunnus. Üks valik: {alahindama; objektiivselt hindama; ülehindama}.
27. Kui tähtis on Teile isiklik suhtlemine ümbritsevate inimestega? (1 – pole üldse tähtis, 5 – väga tähtis)
28. Kui tähtis on Teile suhtlemine sotsiaalvõrkudes? (1 – pole üldse tähtis, 5 – väga tähtis)
29. Teie peamine hobi/ huviala. Nominaaltunnus. Unikaalne tekst.
30. Kas olete osalenud selle hobiga/huvialaga seotud võistlustel/ üritustel? Binaarne tunnus: {jah/ ei}.
31. Kommentaarid. Unikaalne tekst.

Tunnuse nimetuse kujundamine

Edaspidi andmestiku D küsimuste vastusevariantidele oli antud lühinimesid. Näiteks, viiendal küsimusel

*Hinnake oma keelteoskust (*1 – *ei oska üldse,* 5 – *oskan suurepäraselt).*
Palutakse hinnata 5.1 *eesti,* 5.2 *vene,* 5.3 *inglise keeled*

on kolm vastust K5_est, K5_ru ja K5_en.

Lisa 6: Küsitluse eeltöötluse kood

Siin on esitatud küsitluse eeltöötluse koodi osa.

```
temp <- ict
# Tunnuste eemaldamine
temp <- temp[,-c(1,4,5,54)]

newNames <- c("Aeg","AegL","kys_keel",
"Nimi","S_aeg","Maakond","Aeg_kol","Emakeel","K5_est","K5_eng","K5_ru","K6",
"K7_Tv","K7_Meed","K7_An","K7_Prog","K7_Nuti", "K7_ML",
"K7_Side","K7_Andu","K8_opTu","K9_huVa","Vorm","K11_eelH",
"K12_eelN","K13_vNr","K14_TooK", "K15_K14",
"K16_mat","K16_fys","K16_prog","K16_eng","K16_IT","K16_aut","K16_pkav",
"K17_olva","K18_opet", "K19_maine",
"K20_toet","K21_ootus","K22_Huma","K22_Reaal","K23_mskd","K24_roll","K25_rask","K26_ens
h","K27_suht","K28_sotV","K29_hobi","K30_hobiTug")

colnames(temp) <- newNames
```

```
#29

temp$K29_hobi <- ifelse(is.na(temp$K29_hobi),"-",temp$K29_hobi)
temp$K29_hobi <- as.factor(as.character(temp$K29_hobi))

#8
temp$Keel <- ifelse(temp$Keel=="русский","vene",temp$Keel)
levels(temp$Keel) <- c("ru","est")
temp$Keel <- as.factor(temp$Emakeel)

# 3
unique(temp$kys_keel)
#"eesti"   "русский" "Русский"
temp$kys_keel <- as.factor(temp$kys_keel)
levels(temp$kys_keel) <- c("eesti","vene","vene")


# Loodud tulbad
#aeg
vast_min <- temp$Aeg
temp$Aeg <- as.numeric(format(vast_min,"%m"))
temp$Aeg <- ifelse(temp$Aeg<6,"kevad","suvi")
temp$Aeg <- as.factor(temp$Aeg)
#vastamise pikkus
temp$AegL <- temp$AegL-vast_min
temp$AegL <- round(as.double(temp$AegL),1)
#vanus
temp$Vanus <- floor(age_calc(as.Date(temp$S_aeg), units = "years"))
```

Lisa 7: Küsimuste esmaanalüüs

See lisa sisaldab 1-5 skaalaga küsimuste kirjeldavat analüüsi.

```
-- Variable type: numeric ------------------------------------------------------------------
# A tibble: 28 x 11
   skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
 * <chr>             <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1 K5_est                0             1  3.87 0.956     1     3     4     5     5 ▁▂▅▇▇
 2 K5_eng                0             1  3.52 0.867     1     3     4     4     5 ▁▃▇▃▂
 3 K5_ru                 0             1  4.24 1.35      1     4     5     5     5 ▁▁▁▁▇
 4 K6                    0             1  3.51 0.814     2     3     4     4     5 ▂▇▂▅▂
 5 K7_Tv                 0             1  2.54 1.23      1     2     2     3     5 ▇▇▇▂▂
 6 K7_Meed               0             1  2.92 1.23      1     2     3     4     5 ▇▇▇▃▂
 7 K7_An                 0             1  4.38 0.833     2     4     5     5     5 ▁▂▁▃▇
 8 K7_Prog               0             1  4.69 0.667     1     5     5     5     5 ▁▁▁▁▇
 9 K7_Nuti               0             1  4.54 0.854     1     4     5     5     5 ▁▁▁▂▇
10 K7_ML                 0             1  4.10 1.09      1     4     4     5     5 ▁▂▃▅▇
11 K7_Side               0             1  4.72 0.564     2     5     5     5     5 ▁▁▁▂▇
12 K7_Andu               0             1  4.10 1.07      1     3     4     5     5 ▁▂▅▇▇
13 K16_mat               0             1  3.57 0.940     1     3     3     4     5 ▁▂▇▃▂
14 K16_fys               0             1  3.25 0.957     1     3     3     4     5 ▁▃▇▃▂
15 K16_prog              0             1  3.01 1.20      1     2     3     4     5 ▇▇▇▇▂
16 K16_eng               0             1  3.80 0.932     1     3     4     4     5 ▁▂▅▇▂
17 K16_IT                0             1  2.70 1.33      1     2     3     4     5 ▇▇▇▇▂
18 K16_aut               0             1  2.40 1.35      1     1     2     3     5 ▇▇▅▂▂
19 K16_pkav              0             1  2    1.23      1     1     2     3     5 ▇▅▂▁▁
20 K17_olva              0             1  4.56 0.583     3     4     5     5     5 ▁▂▁▃▇
21 K19_maine             0             1  4.49 0.605     3     4     5     5     5 ▁▂▁▃▇
22 K20_toet              0             1  4.73 0.579     3     5     5     5     5 ▁▂▁▂▇
23 K21_ootus             0             1  4.78 0.420     4     5     5     5     5 ▁▂▁▁▇
24 K22_Huma              0             1  3.43 1.01      1     3     3     4     5 ▁▃▇▃▂
25 K22_Reaal             0             1  3.94 0.817     2     3     4     5     5 ▁▂▇▃▇
26 K23_mskd              0             1  4.21 0.776     2     4     4     5     5 ▁▂▁▃▇
27 K27_suht              0             1  4.17 0.843     1     4     4     5     5 ▁▂▁▃▇
28 K28_sotV              0             1  3.45 1.11      1     3     3     4     5 ▂▃▂▇▅
```

Lisa 8: Programmeerimiskeele R paketi car VIF() funktsioon

```{r}
library(car)
getS3method("vif", "default")
```

```
function (mod, ...)
{
    if (any(is.na(coef(mod))))
        stop("there are aliased coefficients in the model")
    v <- vcov(mod)
    assign <- attr(model.matrix(mod), "assign")
    if (names(coefficients(mod)[1]) == "(Intercept)") {
        v <- v[-1, -1]
        assign <- assign[-1]
    }
    else warning("No intercept: vifs may not be sensible.")
    terms <- labels(terms(mod))
    n.terms <- length(terms)
    if (n.terms < 2)
        stop("model contains fewer than 2 terms")
    R <- cov2cor(v)
    detR <- det(R)
    result <- matrix(0, n.terms, 3)
    rownames(result) <- terms
    colnames(result) <- c("GVIF", "Df", "GVIF^(1/(2*Df))")
    for (term in 1:n.terms) {
        subs <- which(assign == term)
        result[term, 1] <- det(as.matrix(R[subs, subs])) * det(as.matrix(R[-subs,
            -subs]))/detR
        result[term, 2] <- length(subs)
    }
    if (all(result[, 2] == 1))
        result <- result[, 1]
    else result[, 3] <- result[, 1]^(1/(2 * result[, 2]))
    result
}
```