TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Kertu-Carina Kallaste

Ülikoolist väljalangemise ennustamine masinõppe mudelite abil

Bakalaureusetöö (9 EAP)

Juhendajad: Leo Siiman, PhD
Elena Sügis, PhD

Tartu 2023

Ülikoolist väljalangemise ennustamine masinõppe mudelite abil

Lühikokkuvõte: Kõrge väljalangevus on aktuaalne probleem kõrgkoolides üle maailma. Kuna teema puudutab negatiivselt nii tudengeid ja kõrgkoole kui ka ühiskonda tervikuna, on väljalangemisriski ennustamine populaarne uurimisvaldkond. Antud bakalaureusetöö eesmärk on luua masinõppe mudel Tartu Ülikooli bakalaureuse-, rakenduskõrghariduse ja integreeritud õppe tudengite väljalangemisriski prognoosimiseks, kasutades selleks Tartu Ülikooli õppeinfosüsteemi poolt aastatel 2011 kuni 2022 kogutud õpianalüütilisi andmeid.

Uurimistöö raames rakendati ja hinnati mitmeid erinevate algoritmiliste lähenemisviisidega masinõppe mudeleid, võimaldamaks nende ulatuslikku võrdlevat analüüsi. Täpsemini saavutati parimad ennustustulemused otsustusmetsa algoritmil põhineva ennustusmudeliga, mis suutis testandmestikul tuvastada 88% väljalangejatest. Mudeli ROC AUC skoor oli 0,94, mis viitab väga kõrgele klasside eristamise võimekusele. Kuigi need tulemused on paljulubavad, on üldistatavuse tagamiseks siiski oluline uuemate andmete lisandumisel kontrollida mudeli toimivust.

Praktilise töö tulemusena loodi riskimudelite kogum, millel on võrreldes praegu Tartu Ülikoolis kasutatava mudeliga parem ennustamisvõime. Tulevikus on võimalik loodud mudelid integreerida ülikooli õpianalüütika töölauaga, mis võimaldaks programmijuhtidel riskiolukordades ennetavalt sekkuda.

Võtmesõnad: masinõpe, otsustusmets, XGBoost, logistiline regressioon, Python, klassifitseerimisalgoritm, tasakaalustamata andmed, ülikool, väljalangeja

CERCS: P176, Tehisintellekt

University dropout prediction using machine learning models

Abstract: High dropout rates are a relevant problem in higher education institutions all over the globe. As this issue negatively affects both students, universities, and society as a whole, predicting dropout risk has become a popular research field. The aim of this bachelor's thesis is to create a machine learning model for predicting the dropout risk of bachelor, applied higher education and integrated study program students in University of Tartu, using the educational analytical data collected by the study information system of University of Tartu from 2011 to 2022.

We have implemented and evaluated several machine learning models, each utilizing distinct algorithmic approaches, to facilitate a comprehensive comparative analysis. More specifically, the best prediction results were achieved with a prediction model based on the random forest algorithm, which was able to identify 88% of dropouts on the test set. The model's ROC AUC score was 0.94, indicating very high ability to distinguish between classes. Although these results are promising, it is important to verify the model's performance once newer data becomes available to ensure its generalisation.

As a result of this practical work, a set of risk models with improved predictive power over the currently deployed model was created. An outcome of this research has the potential to be integrated into the university's educational analytics dashboard in the future. This would allow program managers to proactively intervene in risk situations.

Keywords: machine learning, random forest, XGBoost, logistic regression, Python, classification algorithm, imbalanced data, university, dropout

CERCS: P176, Artificial intelligence

Sisukord

Sissejuhatus

Kõrge ülikoolist väljalangemise määr on aktuaalne probleem kõrgkoolides üle maailma (OECD, 2022). Väljalangemisriski ennustamine on populaarne uurimisvaldkond, kuna teema puudutab negatiivselt nii tudengeid ja kõrgkoole kui ka ühiskonda tervikuna. Lisaks väljalangemisega kaasnevatele individuaalsetele probleemidele on temaatika äärmiselt tähtis kõrghariduse rahastamise küsimuse valguses: oluline on minimeerida ressursside raiskamist kõrgkooliõpinguid pooleli jätvate tudengite koolitamise näol. Kuigi varasemalt on teemal läbi viidud arvukalt uuringuid, ei ole senimaani loodud üldkasutatavat riskimudelit, mis sisaldaks konkreetseid ennustustunnuseid ning oleks universaalselt rakendatav.

Hetkel on Tartu Ülikoolis kasutusel õpianalüütika töölaud, mis kuvab õppejõududele ja programmijuhtidele infot tudengite väljalangemisriskide kohta (Tartu Ülikool, 2022). Kuvatavad ennustused pärinevad otsustusmetsa algoritmil põhinevalt mudelilt, mis on treenitud Tartu Ülikooli (edaspidi ka TÜ) õppeinfosüsteemi kogutud õpianalüütilistel andmetel. Kahjuks on see mudel väljalangemisriski ennustamisel liialt mõõdukas ning tuvastamata jääb enam kui kolmandik riskirühma kuuluvatest tudengitest: 04.11.2022 seisuga oli mudeli saagis (ingl *recall*) bakalaureuse-, rakenduskõrghariduse ja integreeritud õppe tudengite puhul vaid 61 protsenti, magistriõppe puhul veelgi madalam.

Käesoleva bakalaureusetöö eesmärk on luua senisest tulemuslikum masinõppe mudel Tartu Ülikoolist väljalangemise riski prognoosimiseks bakalaureuseõppe tudengite puhul. Väljalangemisriski ennustamisel keskendutakse Tartu Ülikooli bakalaureuse-, rakenduskõrghariduse ja integreeritud õppe tudengitele ning selleks rakendatakse erinevaid masinõppe meetodeid ja algoritme. Loodud mudelite tulemusi võrreldakse olemasoleva õpianalüütika ennustusmudeliga.

Lõputöö on jaotatud kolmeks suuremaks peatükiks. Esimene peatükk annab ülevaate taustinfost: kirjeldatakse töös kasutatavaid põhimõisteid, selgitatakse täpsemalt probleemi olemust ning tutvustatakse varasemaid uurimusi. Lisaks tutvustatakse töös kasutatud masinõppe meetodeid. Teises peatükis kirjeldatakse sammhaaval praktilist tööprotsessi ning eksperimentide ülesehitust. Viimases peatükis antakse ülevaade loodud riskimudelite tulemustest ning antakse nõuandeid edasiste uuringute ja mudeli rakendamise jaoks.

1. Taustinfo

Selles peatükis kirjeldatakse ja selgitatakse töös kasutatavaid põhimõisteid ning masinõppe meetodeid. Ülevaade antakse kõrgkoolist väljalangemisest, selle võimalikest põhjustest ja probleemsetest tagajärgedest nii indiviidile kui ka ühiskonnale, ning varasematest uuringutest. Lisaks sellele kirjeldatakse andmeteaduse valdkonnas levinud CRISP-DM andmeanalüüsi metodoloogiat. Viimasena tutvustatakse juhendatud masinõppe algoritme, *stacking* tehnikat, mudelite headuse mõõdikuid ning levinud andmete tasakaalustamise meetodeid.

1.1. Väljalangemine

Käesoleva bakalaureusetöö raames defineeritakse väljalangejat kui tudengit, kes on vähemalt ühel õppekaval õpingud pooleli jätnud ehk kes on eksmatrikuleeritud mõnel muul põhjusel kui õppekava täies mahus täitmine. Kõrgkoolide jaoks on väljalangejate arvu minimeerimine oluline selleks, et pakkuda tulevastele spetsialistidele maksimaalsel võimalikul tasemel haridust, vältides seejuures ressursside raiskamist (Bargmann jt, 2022). Paraku on kõrge väljalangemismäär aktuaalne probleem kõrgkoolides üle maailma: OECD[1] riikides lõpetab bakalaureuseõppe hiljemalt kolm aastat pärast nominaalaja lõppu keskmiselt 68 protsenti tudengitest (OECD, 2022). See aga tähendab, et pea kolmandikul sisseastujatest jäävad suure tõenäosusega mingil põhjusel õpingud lõpetamata. Olukorra leevendamiseks tuleks võimalikult vara tuvastada potentsiaalselt suurendatud väljalangemisriskiga õpilased ning vajadusel sekkuda. Selleks tuleb alustada väljalangemist soodustavate tegurite välja selgitamisest.

1.2. Õpingute katkestamise põhjused

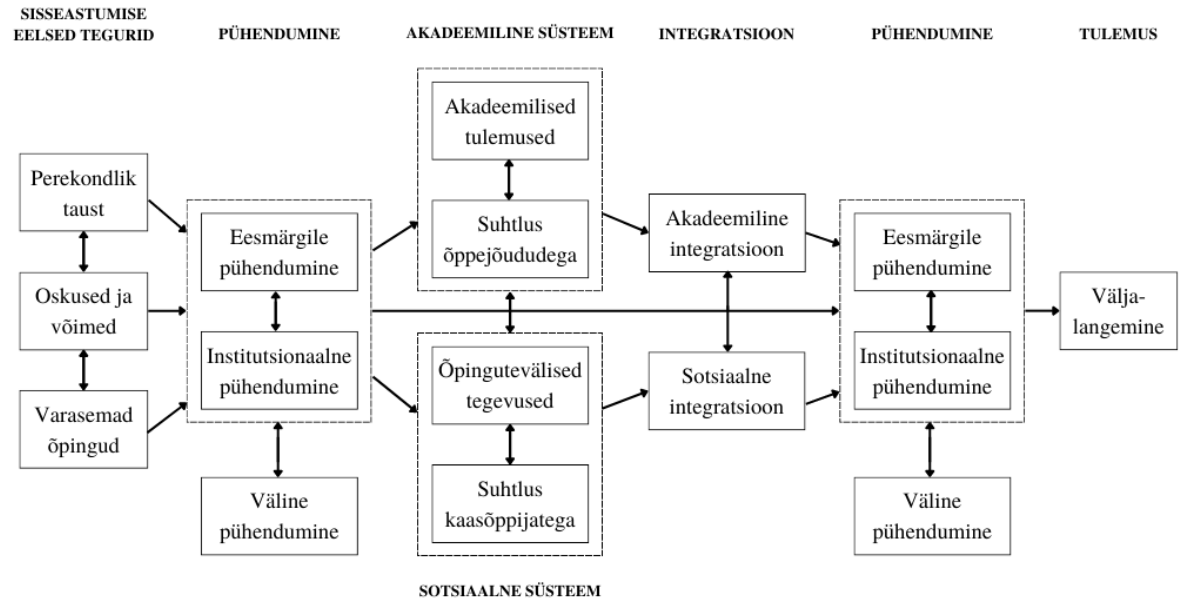
Väljalangemise põhjused on indiviiditi erinevad ning võivad tuleneda nii raskustest õppetöös, ebasobivast erialavalikust kui ka eraelulistest probleemidest. Vincent Tinto, kelle käsitlused on kõrghariduse valdkonnas saavutanud „peaaegu paradigmaatilise staatuse“ (Braxton jt, 2000: 569), on koostanud tunnustatud integratsioonimudeli (Tinto, 1994), mis kirjeldab ja selgitab süvitsi kolme peamist õpingute katkestamise mõjutegurit:

1) õpilase taustaga (kodune keskkond, puudujäägid varasemast haridusteekonnast) seotud raskused kõrgkooliõpingutes;
2) tulevikueesmärgid kõrgkooli astumisel;

[1] Majanduskoostöö ja Arengu Organisatsioon (ingl *Organisation for Economic Co-operation and Development*). https://www.oecd.org/about/

3) akadeemiline ja sotsiaalne integratsioon kõrgkooliõpingute vältel.

Mudelis (vt joonis 1) on väljalangemist visualiseeritud dünaamilise protsessina, mille lõpptulemus kujuneb välja olenevalt tudengi ning haridusasutuse integratsioonist. Vastava teooria kohaselt on tudengitel vaja lõimuda nii akadeemilise kui ka sotsiaalse süsteemiga, et õpingud edukalt lõpetada. Seejuures on ka haridusasutustel oluline roll tudengite jaoks toetava keskkonna loomisel ning lõimumise soodustamisel. Pikisuunalise protsessi vältel omandatud negatiivsed kogemused see-eest vähendavad pühendumist ning soodustavad väljalangemist. Mudelis viitab pühendumine tudengi psühholoogilisele seotusele enda akadeemiliste eesmärkide ning institutsionaalsete väärtustega.



Joonis 1. Tinto integratsioonimudel (Tinto, 1994: 114, kohandatud).

Kuigi mitmed uuringud (Fincham jt, 2021; Karp jt, 2010; Santos-George, 2012) kinnitavad akadeemilise ja sotsiaalse integratsiooni seost väljalangemiskavatsustega, ei ole nende põhjal võimalik kindlaks määrata integratsiooni ja väljalangemise põhjuslikku järjestust. On võimalik, et väljalangemiskalduvusega tudengid ei soovigi integreerumise nimel vaeva näha, mis pöörab põhjusliku järjestuse vastupidiseks (Piepenburg & Beckmann, 2022). Sellest hoolimata on mudelis kirjeldatud mõjutegurid juhtnööriks õpingute katkestamise võimalike põhjuste välja selgitamisel.

1.3. Kaasnevad probleemid ja tagajärjed

Väljalangemine toob Staiculescu jt juhtumiuuringu (2019) põhjal kaasa probleeme nii üksikisiku kui ka ühiskonna tasandil. Individuaalsest vaatepunktist jäävad pooliku haridustee

korral puudulikuks mitmed professionaalsed kompetentsid, mis vähendab produktiivsust. Lisaks sellele võib õpingutes ebaõnnestumine ka psühholoogiliselt raskelt mõjuda, tekitades stressi ning suurendades ebakindlust. Ühiskonna vaatepunktist väheneb väljalangemismäära suurenedes ühiskonna toimimise ja arengu tagamiseks vaja minevate kõrgelt kvalifitseeritud spetsialistide arv. Madalamalt kvalifitseeritud tööjõud on see-eest autorite sõnul suurema tõenäosusega töötu või marginaliseeritud, mistõttu võidakse vajada riigipoolset abi.

Kuigi mitmetes riikides, sealhulgas Eestis, on kõrghariduse rahastamises suurim osa riigil, ei ole võimalik eeldada, et riik suudab kõrgkoolidele tagada piisava rahastuse kvaliteetse hariduse pakkumiseks (Valk jt, 2022). Sellisel juhul on järjepidev ressursside raiskamine õpinguid pooleli jätvate tudengite koolitamise näol rahaliselt koormav kogu ühiskonnale.

1.4. Varasemad uurimused

Kõrgkoolist väljalangemise teemal on läbi viidud võrdlemisi palju uuringuid, kuna väljalangemine mõjutab lisaks üksikisikule ka hariduslike otsuste eest vastutavaid üksuseid: akadeemilisi asutusi ning riiki (Chounta jt, 2020). See peatükk annab ülevaate varasemate uuringute tulemustest, seejuures põhjalikumalt kirjeldatakse masinõppe meetodite abil tehtud avastusi.

Chounta jt (2020) hindasid õpingute poolelijätmise tõenäosust, kasutades arvutuslikku lähenemist. Mudelite treenimiseks ning testimiseks kasutati Tartu Ülikooli õppeinfosüsteemi poolt kogutud andmeid bakalaureuse- ning integreeritud õppe tudengite kohta. Väljalangemisriski ennustamiseks jaotati andmestiku tunnused kolme dimensiooni: tudengi akadeemiline taust, osavõtlikkus ning akadeemilised saavutused. Seejärel loodi iga dimensiooni jaoks eraldi klassifikaator. Kõigi kolme klassifikaatori kombineerimisel saadud ennustusmudel jaotas tudengid nende väljalangemisriski tõsiduse põhjal kolme gruppi: madal, keskmine ja kõrge risk. Mida suurem arv dimensioone väljalangemist tõenäoliseks pidas, seda kõrgema riskiga gruppi tudeng klassifitseeriti. Seos osavõtlikkuse ning akadeemiliste saavutuste klassifikaatorite vahel oli väga tugev[2] (ρ=0,92, p<0,001). Kuigi viimase kahe seos akadeemilise tausta klassifikaatoriga oli väga nõrk (ρ<0,2, p<0,001), saavutati parimad ennustustulemused siiski kõigi kolme dimensiooni kaasamisel: kombineeritud mudeli saagis (ingl *recall*) oli 0,97 ehk testandmestikul suudeti tuvastada suisa 97 protsenti väljalangejatest. Erinevate teaduskondade programmijuhid võtsid mudeli põhjal loodud hoiatussüsteemi küll meeleldi vastu, kuid tõid

[2] Seose tugevuse hindamiseks kasutati Spearmani korrelatsioonikordajat (ρ) ning olulisuse tõenäosust (p).

esile, et hoiatused peaksid ilmtingimata olema õigeaegsed ning jätma piisavalt aega sekkumiseks, et väljalangemist ennetada. Kahjuks on esmakursuslaste väljalangemisriski hindamine kirjeldatud uuringus kasutatud tunnuste tõttu keeruline ülesanne: puudub piisav teave akadeemiliste tulemuste kohta.

Milano Polütehnilise Ülikooli teadlased (Cannistrà jt, 2022) võtsid ülikoolist väljalangemise tõlgendamisele keskenduva kontseptuaalse raamistiku loomisel arvesse nii õppeprotsessi kui ka võimalikult varajaste väljalangemisprognooside vajalikkust. Raamistik põhineb õpilase kui indiviidi haridusteekonnal, mis kujuneb mitmete tegurite, näiteks keskkonna ja geenide mõjul, ning toetub ennustuste tegemisel andmetele õpilase varasemate haridusetappide kohta. Kuigi rohkemate andmete olemasolul on prognoosid täpsemad, on siiski oluline leida tasakaal andmete kogumisele kulutatava ajaga: oluline ei ole üksnes ennustuse õigsus (ingl *accuracy*), vaid ka selle õigeaegsus. Uuringu tulemused tõestavad, et masinõppe ja statistiliste mudelite abil on väljalangemisriski võimalikult varajane prognoosimine reaalselt võimalik. Tänu sellele on pärast riski tuvastamist võimalik isikustatult sekkuda ning luua toetavad juhendamissüsteemid, mis järeleaitamist vajavate tudengite raskusi leevendaksid. Artiklis rõhutati ka seda, et oluline on uurida varajasi õppetulemusi mõjutavaid tegureid: eeldatavasti on väljalangemise põhjused olenevalt õpitud semestrite arvust erinevad, kuid eriti oluline seos väljalangemisega on just esimese semestri õppetulemustel.

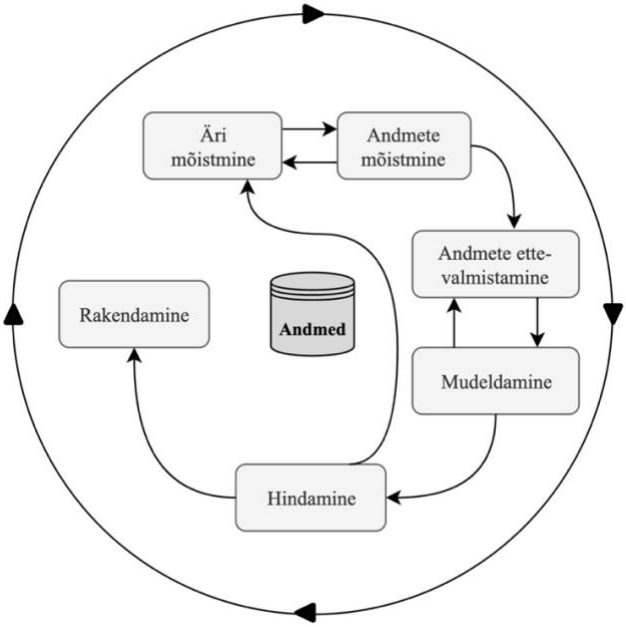
Saksamaal Karlsruhe Tehnoloogiainstituudis (Kemper jt, 2020) ennustati ülikoolist väljalangemist kahe erineva masinõppe lähenemisega: logistilise regressiooni ning otsustuspuudega. Vältimaks tudengite diskrimineerimist õppetöövälise ning potentsiaalselt privaatsusega seotud teabe põhjal, kasutati ennustusmudelite treenimiseks vaid kõrgkoolides vaikimisi salvestatavaid andmeid tudengite õppetulemuste kohta. Kirjeldatava uuringu raames saavutati täpsemad ennustustulemused otsustuspuude abil, kuna nende korral on andmestike ettevalmistusprotsess logistilise regressiooni mudelitega võrreldes vähem kompleksne ning seetõttu vigade tegemise tõenäosus väiksem. Kuna väljalangemist oli võimalik 83-protsendilise õigsusega ennustada juba esimese semestri järel, ilmneb, et tulemuslikkusel on väljalangemisega väga oluline seos. Seega piisab väljalangemisriski hindamiseks ülikoolide kogutavatest õppetulemustega seotud andmetest. Kuigi sarnaselt teistele uuringutele toodi välja, et väljalangevus on kõrgeim esimestel semestritel ning seetõttu prognooside õigeaegsus kriitilise tähtsusega, ei tohi autorite sõnul tähelepanuta jätta ka hilisemate väljalangemisjuhtumite ennustamist.

Mitmeid sarnaseid uuringuid on läbi viidud ka Eesti ülikoolides, ning tulemused on rohkemal või vähemal määral teiste uuringutega kokkulangevad. Näiteks Tallinna Tehnikaülikooli Virumaa Kolledžis läbi viidud juhtumiuuringus (Maksimova jt, 2021) toodi välja, et enam kui 40 protsenti aastatel 2012–2016 Virumaa Kolledžis rakenduslikku arvutiteadust õppinud tudengitest langes välja juba esimesel õppeaastal. Kõrget väljalangejate osakaalu võib allika põhjal seostada tudengitele raskusi valmistavate kursustega, mille ümbertõstmine esimeselt semestrilt teisele ja kolmandale ei toonud oodatud tulemusi väljalangemise määra vähendamise osas. Väljalangemisriski prognoosimiseks kasutati viit erinevat masinõppe algoritmi, mille treenimiseks kasutati lisaks demograafilistele andmetele ka andmeid eelneva ning käesoleva haridusetapi kohta. Seejuures tõusis ennustuste õigsus pärast esimese semestri õppetulemuste lisamist 70 protsendilt 90 protsendile. Kirjeldatud asjaolud viitavad taaskord väljalangemise olulisele seosele õppetulemustega.

Kokkuvõttes ei ole kõrgkoolides väljalangemisriski ennustamiseks vastavalt kirjandusuuringule ja autori teadmistele kasutusel universaalset üldkasutatavat mudelit, mis sisaldaks selgelt määratletud ennustustunnuseid.

1.5. CRISP-DM metodoloogia

CRISP-DM (ingl *Cross-Industry Standard Process For Data Mining*) on andmekaeve ja andmeteaduse projektide läbiviimise metodoloogia (Chapman jt, 2000). Joonisel 2 on kujutatud CRISP-DM protsessimudel, mis annab ülevaate andmeanalüüsi protsessi etappidest.



Joonis 2. CRISP-DM protsessimudel (Chapman jt, 2000: 13, kohandatud).

CRISP-DM protsessi võib jagada kuueks suuremaks etapiks. Enamasti hakatakse etappe lahendama joonisel väljatoodud järjekorras, kuid tihtilugu on vaja nendes sammudes ka tagasi pöörduda. Järgnevalt välja toodud sammud annavad andmeteaduse meetodite rakendamisele formaalse raamistiku ja aitavad seda siduda spetsiifiliste eesmärkidega ning tulevase rakendamisega:

1. **Äri mõistmine** - valdkonna taustinfo põhjal projekti eesmärkide ja nõuete defineerimine, andmeanalüüsi probleemi määratlemine ning esialgne kava püstitatud eesmärkide saavutamiseks.
2. **Andmete mõistmine** - alusandmestikuga tutvumine ning andmete kirjeldus.
3. **Andmete ettevalmistamine** - andmete eeltöötlus, mis võib sisaldada näiteks andmete puhastamist, filtreerimist, tasakaalustamist, ühendamist teiste andmeallikatega, tunnuste sobivale kujule viimist jpm. Etapi lõpptulemuseks on andmestik, mida kasutatakse sisendina mudeldamise etapis.
4. **Mudeldamine** - mudeldamistehnikate ja algoritmide valik, valitud tehnikate rakendamine ning mudeli parameetrite optimeerimine. Mudeli võimsuse ja üldistatavuse hindamine mudeli headuse mõõdikute abil.
5. **Hindamine** - mudeli hindamine projekti raames, et veenduda selle sobivuses lõpliku kasutuselevõtu jaoks, edasiste tegevuste määratlemine.
6. **Rakendamine** - kasutuselevõtukava koostamine, seire- ja hooldusstrateegia planeerimine, tagasivaade ja hinnang projektile ning lõpliku raporti koostamine.

Kirjeldatud tegevussammude järjestus ei ole rangelt määratud ning etappidevahelisi seoseid illustreerib joonis 2.

Kokkuvõttes pakub CRISP-DM metoodika kasutamine struktureeritud raamistikku käesolevas töös püstitatud probleemi lahendamiseks. Lisaks võimaldab samm-sammult lähenemine andmete ja erinevate algoritmide sobivuse põhjalikku uurimist. See aitab paremini tuvastada võimalikke üliõpilaste väljalangemise riskitegureid Tartu Ülikooli andmetes.

1.6. Masinõpe

Selles alapeatükis antakse lühiülevaade töös kasutatud masinõppe meetoditest ja algoritmidest, et toetada arusaama töö tehnilisemast poolest.

1.6.1. Masinõppe algoritmid

Üks selle lõputöö eesmärkidest on katsetada erinevaid masinõppe algoritme, luua nende abil erinevaid klassifitseerimismudeleid ning võrrelda nende tulemusi baasmudeli tulemustega. Klassifitseerimisalgoritmide puhul on tegu juhendatud masinõppe algoritmidega (ingl *supervised machine learning algorithms*): vastavad mudelid õpivad etteantud tunnuste ja märgendite abil, kus märgend on mingi kategooriline väärtus (Gupta jt, 2022).

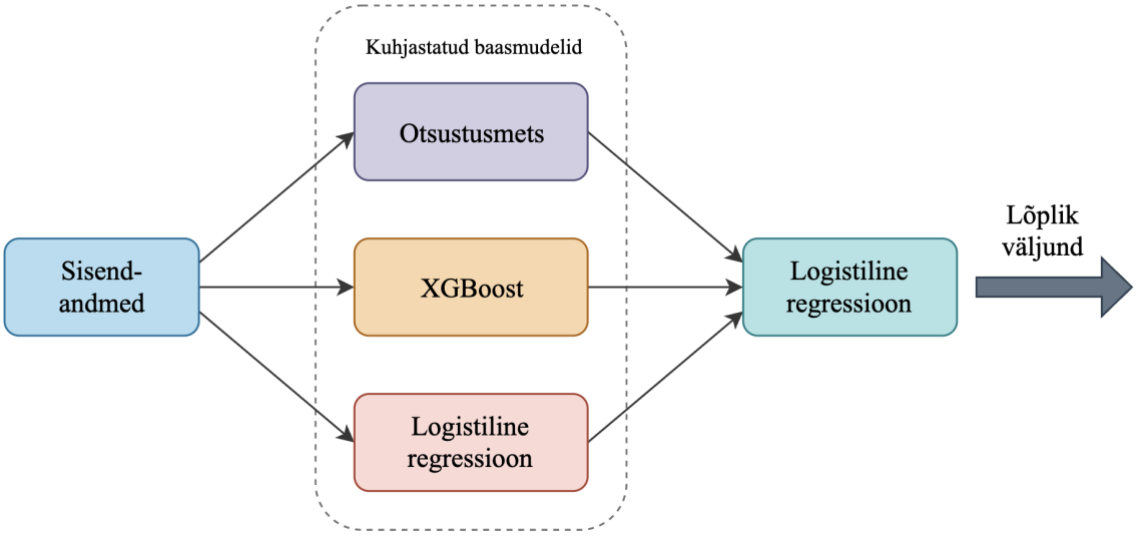
Otsustuspuu (ingl *decision tree*) ning otsustusmets (ingl *random forest*) on laialdaselt levinud klassifitseerimisalgoritmid (Jain jt, 2022; Navada jt, 2011). Järgnev otsustuspuu kirjeldus põhineb Rokachi ja Maimoni kirjutatud õpikul (2007). Nagu algoritmi nimetusest ilmneb, kasutab otsustuspuu otsuste tegemisel puukujulist graafi. Puu sisemised sõlmed on otsustussõlmed, kus tuleb teha valik kõigi võimalike alternatiivide vahel. Terminaalsõlm ehk leht, milleni jõutakse mööda otsustusharusid liikudes, määrab ennustatava klassi. Seejuures on võimalik fikseerida puu sügavus, milleni jõudes tuleb teha lõplik klassifitseerimisotsus, et vältida mudeli ülesobitamist treenimisandmetel. Leo Breiman (2001) töötas välja otsustuspuudel põhineva otsustusmetsa algoritmi: iga puu annab enda poolt hääle ennustatavale klassile ning lõplik otsus tehakse enamushääletuse meetodil. Juhuslikkus treenimisandmete ning otsustussõlmedes kasutatavate tunnuste valimisel aitab mudelil paremini üldistusi teha.

Sarnaselt otsustusmetsale põhineb ka XGBoost (*Extreme Gradient Boosting*, otsetõlkes ekstreemne gradiendiga võimendamine) algoritm otsustuspuudel (Chen & Guestrin, 2016). Autorite kirjelduse põhjal loob kõnealune algoritm ühekaupa otsustuspuid ning iga iteratsiooniga püütakse uue puu loomisel vältida eelmiste puude tehtud klassfitseerimisvigu. Selleks antakse suuremad kaalud ridadele ehk näidetele (ingl *sample*), mida eelnevad puud suutsid korrektselt klassifitseerida ning väiksemad kaalud näidetele, mille klassifitseerimisel eelnevad puud eriti edukad ei olnud. Analoogiliselt otsustusmetsale tehakse lõplikud klassifitseerimisotsused ansambelmeetodil (ingl *ensemble method*): individuaalsete puude ennustustele antakse kaal lähtuvalt nende tulemuslikkusest treeningandmetel.

Binaarsetest klassifitseerimisalgoritmidest on laialt levinud logistiline regressioon (ingl *logistic regression*), mille töötas välja statistik David Cox (1958). Algoritmi tööpõhimõte seisneb andmetele logistilise funktsiooni sobitamises: see funktsioon võimaldab leida tõenäosuse

(lõigus 0–1), et uuritava tunnuse väärtus on 1[3]. Seejärel määratakse igale reale binaarne väärtus olenevalt valitud tõenäosuslävest, mille väärtus on vaikimisi 0,5.

Lisaks individuaalsete algoritmide kasutamisele võib neid omavahel ka kombineerida. Seda on võimalik teha kuhjastamise (ingl *stacking*) meetodil, mille pakkus algselt välja David H. Wolpert (1992). Meetod näeb ette mitme erineva masinõppe algoritmi rakendamist, misjärel agregeeritakse saadud mudelite väljundid ehk ennustused. Lõplike ennustuste saamiseks võib kasutada näiteks *meta-learner* (otsetõlkes meta-õppija) mudelit, mis õpib baasalgoritmide väljunditest: selleks sobib vabalt valitud klassifitseerimis- või regressiooni algoritm (Dey jt, 2021; Rokach, 2010). Joonis 3 näitlikustab eelnevalt kirjeldatud kuhjastamise protsessi olukorras, kus sisendandmete eeltöötlus on juba teostatud.



Joonis 3. Mudelite ansambel või mudelite kuhjastamise meetod (Singh, 2021, kohandatud).

Kuhjastamine on üks levinumatest meta-õppe tehnikatest ning selle rakendamine aitab oluliselt parandada masinõppe algoritmide jõudlust ja efektiivsust, optimeerides õppimisprotsessi ennast (Rokach, 2010).

1.6.2. Mudelite headuse mõõdikud

Klassifitseerimismudelite headust hinnatakse erinevate mõõdikute abil. Levinud mõõdikute hulka kuuluvad näiteks õigsus (ingl *accuracy*), täpsus (ingl *precision*), saagis (ingl *recall*), F1-

[3] Binaarsed tunnused kodeeritakse tavaliselt väärtustega 1 ja 0 vastavalt mingi atribuudi olemasolule või puudumisele vaatlusaluses näites.

skoor ning ROC-kõvera alune pindala (ROC-AUC ehk ingl *area under the receiver operating characteristic curve*, otsetõlkes vastuvõtja tööomaduskõvera alune ala).

Õigsus näitab sisuliselt, kui suur osa mudeli ennustustest olid õiged. Binaarse klassifitseerimise korral kasutatakse õigsuse arvutamiseks järgnevat valemit:

$$Õigsus = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

kus TP tähistab tõelisi positiivseid (ingl *true positives)*, TN tõelisi negatiivseid (ingl *true negatives*), FP valepositiivseid (ingl *false positives*) ja FN valenegatiivseid (ingl *false negatives*). Kahjuks ei ole ainult õigsuse maksimeerimine tasakaalustamata andmestike puhul piisav ("Classification: Accuracy", 2022).

Täpsus näitab mudeli konkreetse klassi ennustuste kvaliteeti: kui suur osa konkreetsesse klassi klassifitseeritud näidetest kuuluvad reaalselt sellesse klassi. See aitab hinnata, kui sageli mudel positiivseid vaatlusi korrektselt ennustab. Binaarse klassifitseerimise puhul kasutatakse täpsuse arvutamiseks järgnevat valemit:

$$Täpsus = \frac{TP}{TP+FP}, \quad (2)$$

kus TP tähistab tõelisi positiivseid ning FP valepositiivseid. Saagis see-eest vastab küsimusele, kui suure osa konkreetsesse klassi kuuluvatest näidetest suudab mudel korrektselt tuvastada. Valem mudeli saagise arvutamiseks on järgnev:

$$Saagis = \frac{TP}{TP+FN}, \quad (3)$$

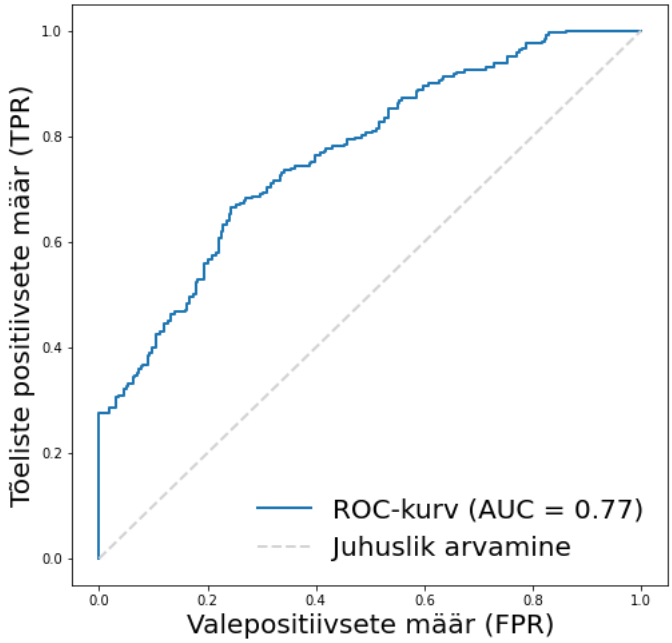
kus TP tähistab tõelisi positiivseid ning FN valenegatiivseid. Nii täpsus kui ka saagis on mudeli headuse hindamisel olulised mõõdikud, kuid kahjuks väheneb enamasti täpsuse suurendamisel saagis ning vastupidi ("Classification: Precision and Recall", 2022). Seega oleneb maksimeeritava mõõdiku valik uuritava probleemi olemusest (Michelucci, 2018).

Sageli on mõistlik mudeli headuse hindamiseks kasutada hoopis F1-skoori, mis aitab leida parima tasakaalu täpsuse ning saagise vahel: F1-skoori maksimeerimisel maksimeeritakse täpsuse ning saagise harmooniline keskmine (Michelucci, 2018). F1-skoori arvutamiseks kasutatakse järgnevat valemit:

$$F1 = 2 \times \frac{täpsus \times saagis}{täpsus + saagis} \quad (4)$$

F1-skoori nimetatakse ka tasakaalustatud F-skooriks ("sklearn.metrics.f1_score", s.a.).

Mudelite headuse hindamiseks võib kasutada ka ROC AUC skoori ("Classification: ROC Curve and AUC", 2022). Skoori väärtus jääb alati lõiku 0–1, kusjuures väärtus 0,5 viitab juhuslikule arvamisele. ROC-kõvera alune pindala esitatakse enamasti graafikuna, millel ROC-kõver näitab mudeli headust kõikide klassifitseerimislävede korral. Joonisel 4 on näide ROC graafikust.



Joonis 4. ROC graafiku näide.

Kõvera parameetriteks on saagis ehk tõeliste positiivsete määr (TPR, ingl *true positive rate*) ning valepositiivsete määr (FPR, ingl *false positive rate*). Mida madalam on klassifitseerimislävi, seda suurem arv näiteid klassifitseeritakse positiivseteks. Seejuures suurenevad nii tõeliste positiivsete kui ka valepositiivsete määrad.

1.6.3. Andmete tasakaalustamise meetodid

Reaalsetes andmestikes esineb sageli klasside tasakaalustamatus, kuid selliste andmete peal õppides kipuvad mudelid näiteid liigitama enamusklassi, jättes vähemusklassi tähelepanuta (Fernández jt, 2018). Probleemi on võimalik lahendada andmete tasakaalustamise abil, milleks on mitu levinud meetodit:

- **Juhuslik ülevalimine (ingl *random oversampling*)** - duplitseeritakse juhuslikult vähemusklassi näiteid (Ali jt, 2019).
- **Juhuslik alavalimine (ingl *random undersampling*)** - eemaldatakse juhuslikult enamusklassi näiteid (Ali jt, 2019).

- **SMOTE (*Synthetic Minority Oversampling Technique*, otsetõlkes sünteetiline vähemuse ülevalimistehnika)** - lisatakse juurde sünteetiliselt genereeritud vähemusklassi näiteid (Chawla jt, 2002).

Kuna alavalimisega kaasneb õppimisprotsessi jaoks kasulike andmete kadu, eelistatakse enamjaolt ülevalimist. Juhusliku ülevalimise korral on aga ülesobitamise oht suurem: genereeritud vähemusklassi eksemplarid on identsed algsete vähemusklassi näidetega. SMOTE aitab seda probleemi vältida ning on tänu oma lihtsusele ka valdkonnas laialdaselt tunnustust leidnud (Ali jt, 2019).

2. Metoodika

Selles peatükis kirjeldatakse tööprotsessi ning eksperimentide ülesehitust. Praktilises osas kasutati peamiste töövahenditena Jupyter Notebooki, mis on interaktiivne veebipõhine arvutuskeskkond (*Project Jupyter*, s.a.), ning Pythoni programmeerimiskeelt (Rossum & Drake, 2009). Pythonil on arvukalt teeke, mida sageli andmeteaduse ning masinõppe valdkondades kasutatakse. Antud töö raames leidsid nende hulgast enim kasutust scikit-learn (Pedregosa jt, 2011), pandas (McKinney, 2010), NumPy (Harris jt, 2020) ning Matplotlib (Hunter, 2007).

Töö ülesehituse kavandamisel lähtuti CRISP-DM metodoloogiast, mis on masinõppe valdkonnas laialdaselt tunnustatud standardprotsess projektide läbiviimiseks.

2.1. Andmete kirjeldus

Lõputöös kasutatakse õpianalüütilisi andmeid TÜ bakalaureuse-, rakenduskõrghariduse ning integreeritud õppe tudengite kohta, mis on kogutud Tartu Ülikooli õppeinfosüsteemi poolt aastatel 2011–2022.

Alusandmestik koosneb 46 erinevast tunnusest ning sisaldab teavet nii üliõpilaste akadeemilise tausta, tulemuslikkuse kui ka pingutuse kohta. Nende hulgas on 38 tunnust numbrilised ja ülejäänud 8 kategoorilised. Kokku sisaldab alusandmestik 48 524 näidet, millest 30 404 vastavad bakalaureusetaseme, rakenduskõrghariduse ning integreeritud õppe tudengitele.

Ennustatav tunnus ’dropout’ (väljalangeja) on binaarne ning selle väärtus on 1 või 0 vastavalt sellele, kas tudeng langes välja ja jättis antud õppekaval õpingud pooleli või mitte. Alusandmestik on võrdlemisi kaldus mitte-väljalangenute suunas: 48 524 näitest on vaid 8133 juhul tegu väljalangenud tudengiga.

2.2. Andmete ettevalmistamine

Et andmestik masinõppe algoritmide jaoks puhastada, viidi läbi mitmeid transformatsioone:

1. Hetkeseisuga TÜ õpilaste nimekirjas olevatele tudengitele vastavate ridade eemaldamine. Mudelite treenimiseks ja testimiseks on võimalik kasutada vaid selliseid näiteid, mille puhul on teada väljalangemise kohta infot sisaldav märgend (’dropout’).
2. Teadaolevalt puuduvate väärtuste täitmine. Olenevalt tunnusest täideti puuduv väärtus kas 0 või õppekava/teaduskonna keskmise väärtusega.

3. Uue tunnuse lisamine tudengi varasemate lõpetamata Tartu Ülikooli õpingute jaoks. Selleks lahutati iga tudengi varasemate TÜ õpingute arvust lõpetatud õpingute arv.
4. Enne esimese semestri algust eksmatrikuleeritud tudengitele vastavate ridade eemaldamine.
5. Kategooriliste tunnuste kodeerimine (ingl *one-hot encoding*). Kuna enamik masinõppe algoritme töötavad vaid arvandmetega, siis teisendati kõik kategoorilised tunnused arvulisteks.
6. Erinevate identifikaatorite eemaldamine. Andmestik sisaldas mitmeid unikaalseid kategoorilist tüüpi identifikaatoreid, mida ei olnud mõistlik kodeerida ega mudelite treenimisel kasutada.
7. Andmestiku juhuslik jaotamine treenimis- ja testandmestikuks (80% ja 20%).

Kuna andmestik oli üsnagi kaldus, siis viimase ettevalmistava sammuna prooviti erinevaid andmete tasakaalustamise tehnikaid. Tulemusi võrreldes otsustati viimastes eksperimentides SMOTE kasuks, mis oli ka kirjanduse põhjal efektiivseim tehnika (Ali jt, 2019). Valitud tehnikat rakendati ainult treenimisandmete peal, et säilitada testandmestikus andmete tegelik jaotus. Selline lähenemine ennetab ülesobitamist ehk aitab tagada, et hinnang mudelile ei oleks liiga optimistlik.

2.3. Mudeldamine

Lõputöö käigus viidi läbi viis erinevat eksperimenti:

1. Esimese eksperimendi raames võrreldi viie individuaalse mudeli tulemusi. Valitud mudeliteks olid otsustuspuu, otsustusmets, KNN ehk k-lähimat naabrit, GaussianNB ehk Gaussi naiivne Bayes ning AdaBoost ehk adaptiivne võimendus (ingl *adaptive boosting*). Kasutati kolme erineva töötlusega andmestikku: rakendati nii juhuslikku üle- ja alavalimist kui ka andmestiku tasakaalustamata jätmist. Viimase sammuna enne mudelite treenimist viidi läbi andmestiku jaotamine treenimis-, valideerimis- ning testandmestikuks (80%, 10% ja 10%).
2. Teine eksperiment oli ülesehituselt üldjoontes samasugune nagu esimene, kuid juhuslikku üle- ning alavalimist rakendati pärast andmete jaotamist ning ainult treenimisandmestikul.

3. Kolmanda eksperimendi ülesehitus jäi üldjoontes eelmise eksperimendiga samaks, kuid andmestikku jäeti alles vaid 10 olulisemat tunnust ja identifikaator. Olulisemate tunnuste tuvastamiseks kasutati II eksperimendist pärit otsustusmetsa mudelit ja selle 'feature_importances_' atribuuti.
4. Neljanda eksperimendi raames jaotati andmestik treenimis- ning testandmestikuks, misjärel rakendati treenimisandmete tasakaalustamiseks SMOTE tehnikat. Seejärel treeniti kolm erinevat klassifikaatorit: otsustusmets, logistiline regressioon ning XGBoost. Lõplike ennustuste saamiseks rakendati mudelite ansambel: mudelite ennustatud tõenäosused[4] keskmistati ning näide klassifitseeriti väljalangejaks, kui saadud keskmine oli suurem kui 0,5.
5. Viienda ehk viimase eksperimendi raames kasutati ansambliõpet kuhjastamistehnikaga. Mudelite kuhjastamise ülesehitus oli järgmine: esimese taseme mudelitena rakendati otsustusmetsa, logistilist regressiooni ja XGBoosti ning meta-klassifitseerijana kasutati logistilist regressiooni. Andmestik jaotati treenimis- ja testandmestikuks ning treenimisandmetel rakendati SMOTE tehnikat. Seejärel loodi kolm mudelit: otsustusmets, logistiline regressioon ning XGBoost. Et optimeerida mudelite jõudlust, kasutati rist-valideerimise tehnikat ja rakendati parimate hüperparameetrite leidmiseks võrguotsingut (ingl *grid search*). Nagu ilmestab joonis 3, kasutati seejärel lõplike ennustuste saamiseks optimeeritud individuaalsete mudelite väljundeid sisendina meta-klassifikaatorile.

Iga ülaltoodud eksperimendi raames hinnati ka mudelite jõudlust.

2.4. Mudelite hindamine

Kuna töös kasutatud andmestik on võrdlemisi tasakaalustamata ning seega õigsus mudeli headuse hindamiseks vähem sobilik, kasutati peamise mõõdikuna F1-skoori. Sellest hoolimata on saagis käsitletava uurimisprobleemi korral äärmiselt oluline: väljalangemise minimeerimiseks on tähtis võimalikult paljud väljalangemisriskiga tudengid tuvastada, et oleks võimalik ennetavalt sekkuda. Lisaks eelnimetatud mõõdikutele kasutati ka ROC AUC skoori, mis on laialdaselt levinud klassifitseerijate headuse hindamiseks (Brownlee, 2020).

[4] 'predict_proba()' meetod scikit-learn teegist võimaldab leida positiivsesse klassi kuulumise tõenäosused ehk tõenäosused, et tegu on väljalangejaga.

Mudelite jõudlust hinnati igas katses testandmetel ning võrreldi kahe baasmudeliga: Tartu Ülikooli õpianalüütika töölauas kasutatava otsustusmetsal põhineva mudeliga ning samuti TÜ teadlaste poolt arendatud kombineeritud ennustusmudeliga (Chounta jt, 2020).

3. Tulemused ja arutelu

Uurimuse raames katsetati arvukalt masinõppe algoritme kombineeritult erinevate andmete tasakaalustamise tehnikatega. Kogu kood on üles laetud avalikku GitHubi repositooriumisse[5]. Tabelis 1 on välja toodud eksperimentide parimad mudelid koos sisendandmete tasakaalustamisel rakendatud tehnikatega.

Tabel 1. Eksperimentide parimad klassifitseerimismudelid.

	Andmete tasakaalustamine	**Algoritm**
Baasmudel 1	–	otsustusmets
Baasmudel 2	ei ole teada	logistiline regressioon
Eksperiment 1	juhuslik ülevalimine kogu andmestikul	otsustuspuu
Eksperiment 2	juhuslik alavalimine treenimisandmetel	otsustusmets
Eksperiment 3	juhuslik alavalimine treenimisandmetel	otsustusmets
Eksperiment 4	SMOTE treenimisandmetel	mudelite ansambel (otsustusmets + XGBoost + logistiline regressioon)
Eksperiment 5	SMOTE treenimisandmetel	mudelite ansambel, kus esimese taseme mudelid otsustusmets, XGBoost ja logistiline regressioon, meta-õppijana logistiline regressioon
Eksperiment 5.1	SMOTE treenimisandmetel	otsustusmets
Eksperiment 5.2	SMOTE treenimisandmetel	XGBoost
Eksperiment 5.3	SMOTE treenimisandmetel	logistiline regressioon

Ülal toodud ennustusmudelite tulemused on varieeruvad. Tabelis 2 on esitatud teostatud eksperimentide parimate mudelite tulemused väljalangemisriski ennustamisel bakalaureuse-, rakenduskõrghariduse ja integreeritud õppe tudengite seas. Halli värviga on välja toodud olulised tulemused. Võrdluseks on välja toodud ka TÜ õpianalüütika töölauas kasutatava baasmudeli tulemused samadel andmetel ning Chounta jt (2020) loodud mudeli ennustustulemused 2014. aastal immatrikuleeritud tudengite andmetel. Selle mudeli treenimiseks kasutati aastatel 2010–2013 immatrikuleeritud tudengite andmeid.

[5] Repositoorium asub veebiaadressil https://github.com/kertucarina/UniversityDropoutPrediction

Tabel 2. Klassifikaatorite tulemused väljalangemisriski ennustamisel.

	Õigsus	**Täpsus**	**Saagis**	**F1-skoor**	**ROC AUC**
Baasmudel 1	0,83	0,72	0,61	0,66	0,76
Baasmudel 2	0,96	0,95	0,97	0,96	–
Eksperiment 1	0,93	0,90	0,96	0,93	0,93
Eksperiment 2	0,83	0,64	0,92	0,76	0,93
Eksperiment 3	0,83	0,64	0,91	0,75	0,91
Eksperiment 4	0,84	0,66	0,89	0,76	0,93
Eksperiment 5	0,86	0,78	0,73	0,75	0,94
Eksperiment 5.1	0,84	0,68	0,88	0,77	0,94
Eksperiment 5.2	0,86	0,75	0,80	0,77	0,94
Eksperiment 5.3	0,84	0,70	0,79	0,74	0,91

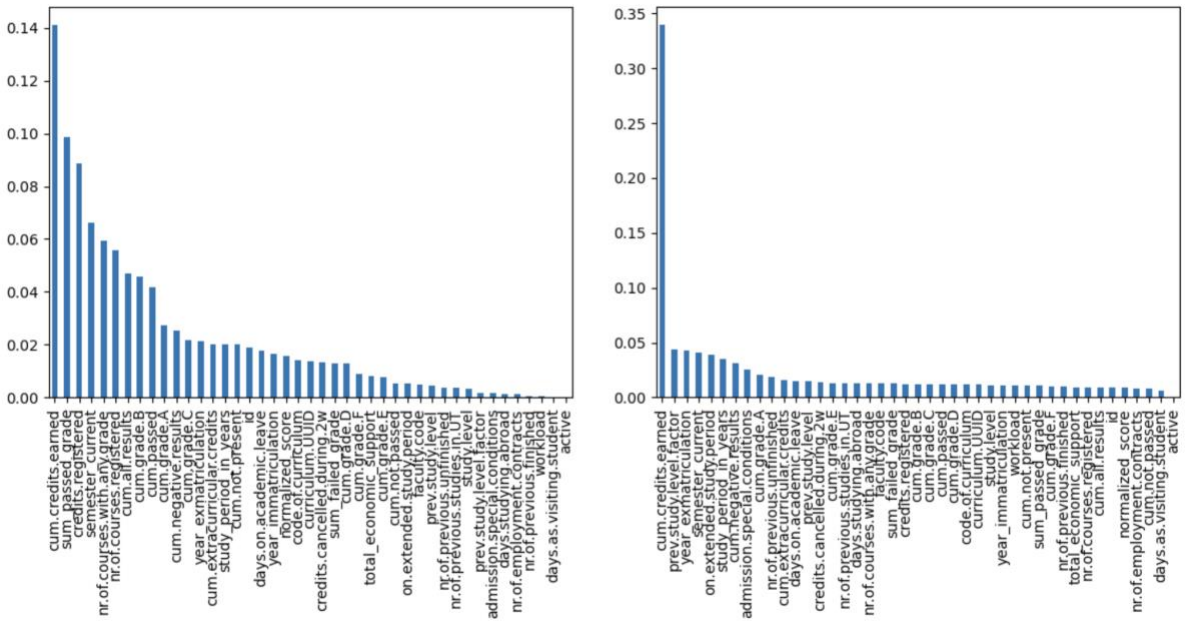
Parimate mudelite valikul lähtuti F1-skoorist ehk täpsuse ja saagise harmoonilisest keskmisest ning ROC AUC skoorist, mis hindab mudeli võimet klasse eristada. Kuigi hinnang esimese eksperimendi mudelile on esmapilgul esimese baasmudeliga võrreldes väga hea, ei ole see asjakohane, kuna ülevalimine teostati kogu andmestikul. See tähendab, et testandmetelt lekkis infot mudeli treenimisprotsessi, mistõttu on hinnang mudelile liialt optimistlik.

ROC AUC skoori järgi saavutati parimad ennustustulemused viienda eksperimendi käigus. Nii individuaalsete mudelite kui ka mudelite ansambli headuse mõõdikud näitavad oluliselt parendatud mudeli ennustusjõudlust võrreldes esimese baasmudeliga. Samuti peegeldavad tulemused, et treenimisandmete tasakaalustamiseks kasutatud SMOTE ehk sünteetiline vähemuse ülevalimistehnika võimaldab suurendada mudeli jõudlust, parandades vähemusklassi ennustamist. Kui andmekogumis on mõni klass ülekaalus ning mudel eelistab näidete liigitamist nendesse enamusklassidesse, aitab SMOTE probleemi leevendada.

Võttes arvesse kõiki mõõdikuid, saavutati parimad ennustustulemused eksperimentide 5, 5.1 ning 5.2 käigus. ROC AUC skoor oli nendes eksperimentides 0,94, mis näitab, et mudelid suudavad väga edukalt klasse eristada. Eksperimendi 5.1 raames loodud otsustusmetsa algoritmil põhineva ennustusmudeli saagis oli 0,88 ehk mudel suutis testandmestikul tuvastada 88 protsenti väljalangenud tudengitest. See-eest on selle mudeli täpsus (0,68) viienda eksperimendi raames loodud mudelite seast kõige madalam. See tähendab, et suur osa tehtud ennustustest on valepositiivsed ehk 32% väljalangejaks klassifitseeritud tudengitest ei olnud tegelikkuses väljalangejad. Kokkuvõtvalt suurendab madalam täpsus programmijuhtide töömahtu, kuid kõrgem saagis võimaldab rohkematel juhtudel väljalangemise ennetamiseks sekkuda.

Kui kõrgem valepositiivsete määr ei ole kriitiline, võivad väljalangemisriski ennustamiseks sobida ka eksperimentide 2, 3 ning 4 raames loodud mudelid, mis suutsid testandmetel tuvastada ligikaudu 90% väljalangenud tudengitest. Võrreldes viienda eksperimendi mudelitega, on nende mudelite täpsused küll märgatavalt madalamad, kuid sellegipoolest on F1-skoorid peaaegu võrdsed.

Eksperimentide 5, 5.1 ning 5.2 raames loodud mudelite õigsused, F1-skoorid ning ROC AUC skoorid on ligilähedased ning individuaalsete mudelite (5.1 ja 5.2) saagised isegi kõrgemad. Kuna individuaalsete mudelite implementatsioon on vähem keerukas, on mudeli seletatavuse lihtsustamiseks mõistlikum kasutusele võtta kumbki nimetatud individuaalsetest mudelitest: optimeeritud hüperparameetritega otsustusmets või XGBoost. Joonisel 5 on välja toodud ennustamisel kasutatud tunnuste olulisus kummagi algoritmi jaoks. Vasakul on esitatud tunnuste olulisused otsustusmetsa jaoks, paremal XGBoosti jaoks. Lisas I tabelis 1 on eesti keeles lahti selgitatud kõikide tunnuste tähendused.



Joonis 5. Tunnuste olulisused otsustusmetsa ja XGBoosti algoritmide jaoks.

Nagu joonistelt näha, aitab mõlema algoritmi korral kõige paremini sihtmuutujat ennustada tunnus ’cum.credits.earned’ ehk läbitud õppeainete maht EAP-des. Lisaks eelnimetatud tunnusele olid mõlema algoritmi jaoks kümne olulisima tunnuse seas ka ’semester_current’ (õpitud semestrite arv), cum.grade.A (hindele A läbitud õppeainete arv) ning ’cum.negative.results’ (negatiivsete tulemuste koguarv).

Kuigi baasmudel 2 on tabeli 2 järgi parim klassifikaator, on sellel mitmeid piiranguid. Esiteks, sarnaselt käesolevale uurimistööle kasutati TÜ õppeinfosüsteemi kogutud andmeid, kuid ei ole midagi teada selle kohta, kas ning kuidas andmeid tasakaalustati. Lisaks olid mudeli treenimiseks ja testimiseks kasutatud andmed käesoleva uurimistööga võrreldes ligikaudu kümme aastat vanemad. On võimalik, et vahepealsel perioodil on andmetes toimunud märkimisväärsed muutused, mis väljalangemisriski ennustamist raskendavad.

Kuna antud lõputöö raames keskenduti ennustusmudelite loomisel bakalaureuse-, rakendus-kõrghariduse ja integreeritud õppe tudengitele, oleks tulevikus kasulik kontrollida, kas loodud mudelid suudavad korrektselt ennustada ka magistritudengite väljalangemisriski. On võimalik, et eri õppetasemete jaoks on vaja ennustusmudeleid kohandada. Uuemate andmete lisandumisel tuleks kindlasti kontrollida ka mudelite headuse hinnangute vastavust tegelikkusele.

2022. aasta sügisel viidi läbi TÜ õpianalüütika töölaua testimine. Kogutud tagasiside[6] näitas, et kasutatav mudel liigitab riskigruppi eeskujulike õppetulemustega tudengeid, kes on akadeemilisel puhkusel pidanud viibima ajateenistuse läbimise tõttu. Seega võiks võimaluse korral juba andmete kogumisel eristada akadeemilisel puhkusel viibimise põhjuseid. Vastava tunnuse lisamine treeningandmestikku võib potentsiaalselt mõjutada mudeli ennustusvõimet.

[6] Töölaua testimise tagasiside dokumenti saab jagada nõudel.

Kokkuvõte

Selle bakalaureusetöö peamine eesmärk oli luua masinõppe mudel Tartu Ülikoolist väljalangemise riski prognoosimiseks bakalaureuse-, rakenduskõrghariduse ning integreeritud õppe tudengite puhul. Kuigi Tartu Ülikoolis on juba kasutusel õpianalüütika töölaud, mis kuvab õppejõududele ja programmijuhtidele infot tudengite väljalangemisriskide kohta, on sealne mudel väljalangemisriski ennustamisel liialt mõõdukas. Seega on senimaani jäänud suur osa väljalangemisriskiga tudengitest tuvastamata.

Mudelite loomisel kasutatud õpianalüütilised andmed on kogutud Tartu Ülikooli õppeinfosüsteemi poolt aastatel 2011 kuni 2022. Töö käigus viidi läbi viis erinevat eksperimenti, mille raames prooviti erinevaid andmete tasakaalustamise meetodeid kombineeritult mitmete klassifitseerimisalgoritmidega: otsustuspuu, otsustusmets, KNN, GaussianNB, AdaBoost, XGBoost ning logistiline regressioon. Enne mudeldamist puhastati alusandmestik puuduvatest väärtustest ning kodeeriti kõik kategoorilised tunnused. Seejärel rakendati nii individuaalseid algoritme kui ka mudelite kombineerimist ansambelmeetodil.

Praktilise töö tulemusena loodi riskimudelite kogum, millel on võrreldes praegu Tartu Ülikoolis kasutatava mudeliga parem ennustamisvõime. Täpsemini, parimad ennustustulemused saavutati otsustusmetsa algoritmil põhineva mudeli abil. Treenimisandmete tasakaalustamiseks rakendati SMOTE tehnikat ning parimate hüperparameetrite tuvastamiseks võrguotsingut. Selle mudeli ROC AUC skoor oli 0,94, mis viitab väga kõrgele klasside eristamise võimekusele. Mudeli saagis oli 0,88 ehk testandmestikul suudeti tuvastada suisa 88% väljalangejatest. Võrdluseks, Tartu Ülikooli õpianalüütika töölauas kasutatava mudeli puhul oli vastav väärtus vaid 0,61.

Kuigi kirjeldatud tulemused on paljulubavad, on siiski oluline uuemate andmete lisandumisel verifitseerida hinnangud mudelite headusele. Tulevikus on võimalik loodud mudelid integreerida ülikooli õpianalüütika töölauaga, võimaldamaks programmijuhtidel ennetavalt sekkuda olukordades, kus tudeng on mingitel põhjustel väljalangemisohtu sattunud.

Viidatud kirjandus

Ali, H., Salleh, M. N. M., Hussain, K., Ahmad, A., Ullah, A., Muhammad, A., Naseem, R., & Khan, M. (2019). A review on data preprocessing methods for class imbalance problem. *International Journal of Engineering & Technology*, *8*, 390–397.

Bargmann, C., Thiele, L., & Kauffeld, S. (2022). Motivation matters: Predicting students' career decidedness and intention to drop out after the first year in higher education. *Higher Education*, *83*. https://doi.org/10.1007/s10734-021-00707-6

Braxton, J. M., Milem, J. F., & Sullivan, A. S. (2000). The Influence of Active Learning on the College Student Departure Process: Toward a Revision of Tinto's Theory. *The Journal of Higher Education*, *71*(5), 569–590. https://doi.org/10.1080/00221546.2000.11778853

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery.

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., & Paganoni, A. M. (2022). Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, *47*(9), 1935–1956. https://doi.org/10.1080/03075079.2021.2018415

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chounta, I.-A., Uiboleht, K., Roosimäe, K., Pedaste, M., & Valk, A. (2020). From Data to Intervention: Predicting Students At-Risk in a Higher Education Institution. Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20): 10th International Conference on Learning Analytics & Knowledge (LAK20). Ed. Kovanović, V., Scheffel, M., Pinkwart, N., & Verbert, K. Society for Learning Analytics Research (SoLAR), 750−755.

Classification: Accuracy. (2022). Google Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy (25.03.2023)

Classification: Precision and Recall. (2022). Google Developers. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (25.03.2023)

Classification: ROC Curve and AUC. (2022). Google Developers. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (25.03.2023)

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

Dey, N., Das, N., & Chaki, J. (2021). *Digital Future of Healthcare*. CRC Press.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.

Fincham, E., Rozemberczki, B., Kovanovic, V., Joksimovic, S., Jovanovic, J., & Gasevic, D. (2021). Persistence and Performance in Co-Enrollment Network Embeddings: An Empirical Validation of Tinto's Student Integration Model. *IEEE Transactions on Learning Technologies*, *14*(1), 106–121. https://doi.org/10.1109/TLT.2021.3059362

Gupta, V., Mishra, V. K., Singhal, P., & Kumar, A. (2022). An Overview of Supervised Machine Learning Algorithm. *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 87–92. https://doi.org/10.1109/SMART55829.2022.10047618

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jain, S., Pandey, K., Jain, P., & Seng, K. P. (2022). *Artificial Intelligence, Machine Learning, and Mental Health in Pandemics: A Computational Approach*. Academic Press.

Karp, M. M., Hughes, K. L., & O'Gara, L. (2010). An Exploration of Tinto's Integration Framework for Community College Students. *Journal of College Student Retention: Research, Theory & Practice*, *12*(1), 69–86. https://doi.org/10.2190/CS.12.1.e

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Maksimova, N., Pentel, A., & Dunajeva, O. (2021). Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study. In M. E. Auer & T. Rüütmann (Eds.), *Educating Engineers for Future Industrial Revolutions* (Vol. 1329, pp. 719–726). Springer International Publishing. https://doi.org/10.1007/978-3-030-68201-9_70

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Michelucci, U. (2018). *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks*. Apress.

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. *2011 IEEE Control and System Graduate Research Colloquium*, 37–42. https://doi.org/10.1109/ICSGRC.2011.5991826

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

Tartu Ülikool. (2022). Õpianalüütika - Learning Analytics. *TÜ Wiki.* https://wiki.ut.ee/pages/viewpage.action?pageId=137335554 (23.03.2023)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Piepenburg, J. G., & Beckmann, J. (2022). The relevance of social and academic integration for students' dropout decisions. Evidence from a factorial survey in Germany. *European Journal of Higher Education*, *12*(3), 255–276. https://doi.org/10.1080/21568235.2021.1930089

Project Jupyter. (s.a.). https://jupyter.org (23.03.2023)

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*. World Scientific.

Rokach, L., & Maimon, O. (2007). *Data Mining with Decision Trees: Theory and Applications* (Vol. 69). WORLD SCIENTIFIC. https://doi.org/10.1142/6604

Rossum, G. V., & Drake, F. L. (2009). *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. CreateSpace Independent Publishing Platform.

Santos-George, A. A. (2012). An Empirical Test of Tinto's Integration Framework for Community Colleges Using Structural Equation Modeling. In *ProQuest LLC*. ProQuest LLC.

Singh, H. (30.03.2021). Advanced Ensemble Learning technique –Stacking and its Variants. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/03/advanced-ensemble-learning-technique-stacking-and-its-variants/

sklearn.metrics.f1_score. (s.a.). Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html (25.03.2023)

Staiculescu, C., Lacatus, M. L., & Richiteanu, N. E.-R. (2019). *Causes And Effects Of University Dropout: Case Study*. 832–838. https://doi.org/10.15405/epsbs.2019.08.03.99

Tinto, V. (1994). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226922461.001.0001

Valk, A., Silm, G., & Tiitsaar, K. (24.04.2022). Ülevaade. Kõrghariduse rahastuskriis jätab palju noori ülikooli ukse taha. *Eesti Rahvusringhääling*. https://www.err.ee/1608575377/ulevaade-korghariduse-rahastuskriis-jatab-palju-noori-ulikooli-ukse-taha

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Lisad

I. Tunnuste kirjeldus

Tabel 1. Ennustamisel kasutatud tunnuste sõnastik.

Tunnus	**Kirjeldus**
active	kas õpib praegu antud õppekaval
admission.special.conditions	kas on vastu võetud eritingimustel
code.of.curriculum	õppekava kood
credits.cancelled.during.2w	summaarne kahe nädala jooksul tühistatud EAP-de arv
credits.registered	summaarne registreeritud EAP-de arv
cum.all.results	tulemuste koguarv
cum.credits.earned	läbitud õppeainete maht EAP-des
cum.extracurricular.credits	läbitud õppekavaväliste õppeainete maht EAP-des
cum.grade.A	hindele A läbitud õppeainete arv
cum.grade.B	hindele B läbitud õppeainete arv
cum.grade.C	hindele C läbitud õppeainete arv
cum.grade.D	hindele D läbitud õppeainete arv
cum.grade.E	hindele E läbitud õppeainete arv
cum.grade.F	hindele F läbitud õppeainete arv
cum.negative.results	negatiivsete tulemuste koguarv (hinne F või ’mittearvestatud’ või ’mitteilmunud’)
cum.not.passed	tulemusega 'mittearvestatud' õppeainete arv
cum.not.present	tulemusega 'mitteilmunud' õppeainete arv
cum.passed	tulemusega 'arvestatud' õppeainete arv
curriculum.UUID	õppekava UUID
days.as.visiting.student	külalisüliõpilasena õpitud päevade arv
days.on.academic.leave	akadeemilisel puhkusel oldud päevade arv
days.studying.abroad	välismaal õpitud päevade arv
dropout	kas on välja kukkunud
faculty.code	valdkonna kood
normalized_score	vastuvõtutulemus
nr.of.courses.registered	summaarne registreeritud kursuste arv
nr.of.courses.with.any.grade	mis tahes hindega kursuste koguarv
nr.of.employment.contracts	summaarne TÜ töölepingute arv

nr.of.previous.finished	edukalt lõpetatud õpingute arv TÜ-s
nr.of.previous.studies.in.UT	varasemate õpingute arv TÜ-s
nr.of.previous.unfinished	varasemate pooleli jäänud õpingute arv TÜ-s
on.extended.study.period	kas on õpinguid pikendanud
prev.study.level	eelmine õppeaste
prev.study.level.factor	eelmise õppeaste faktor
semester_current	õpitud semestrite arv
study_period_in_years	antud õppekohal õpitud aastate arv
study.level	õppeaste
sum_failed_grade	negatiivse hindega kursuste koguarv (hinne F või ’mittearvestatud)
sum_passed_grade	positiivse hindega kursuste koguarv (hinne A-E või ’arvestatud’)
total_economic_support	saadud õppetoetuste kogusumma
workload	õppekoormus
year_exmatriculation	eksmatrikuleerimise aasta
year_immatriculation	immatrikuleerimise aasta

II. Litsents

Lihtlitsents lõputöö reprodutseerimise ja üldsusele kättesaadavaks tegemise kohta

Mina, Kertu-Carina Kallaste,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Ülikoolist väljalangemise ennustamine masinõppe mudelite abil“, mille juhendajad on Leo Siiman ja Elena Sügis, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

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

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

Kertu-Carina Kallaste
09.05.2023