TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Hannes Kuslap

Masinõppe algoritmide rakendamine füüsikast lähtuvate fotomeetriliste punanihete täpsustamisel

Magistritöö (15 EAP)

Juhendaja(d): Elmo Tempel, PhD
Taavi Tuvikene, PhD

Tartu 2025

Masinõppe algoritmide rakendamine füüsikast lähtuvate fotomeetriliste punanihete täpsustamisel

Lühikokkuvõte:

Töös käsitlen masinõppemeetodite rakendamist fotomeetriliste punanihete (*photo-z*) täpsustamiseks, keskendudes füüsikalise mudeli TOPz väljundite parandamisele. Kasutades WAVES uuringu andmestikku, treenin erinevaid regressioonipõhiseid masinõppemudeleid TOPz väljundite reprodutseerimiseks ja täpsustamiseks. Parimaks osutus XGBoost algoritm, mille optimaalne konfiguratsioon suutis oluliselt vähendada ennustusvigu ja erindite osakaalu võrreldes TOPz mudeliga. Seejärel loodi kaks täiendusmudelit: esimene ennustas otseselt punanihke logaritmilist teisendust $\zeta = \ln(1 + z)$, teine aga kogu ζ tõenäosusjaotust. Kombineerisin mõlemad mudelid TOPz väljunditega, kusjuures otse ζ väärtust ennustav mudel saavutas väiksemad vead (MAE = 0,0265) ning tõenäosusjaotuste mudel pakkus suuremat interpreteeritavust ja suutlikkust käsitleda mitmemõttelisi lahendusi. Parimad tulemused saavutati mudelite lineaarse või geomeetrilise kombineerimisega, optimeerides TOPz ja XGBoosti kaalutegurit. Töö näitab, et füüsikalise modelleerimise ja masinõppe sümbioos võimaldab oluliselt täpsemaid ning usaldusväärsemaid fotomeetrilisi punanihke hinnanguid, mis on kriitilise tähtsusega suurte astronoomiliste uuringute jaoks, nagu J-PAS ja WAVES.

Võtmesõnad:
TOPz, fotomeetriline punanihe, WAVES andmestik, XGBoost, masinõpe

CERCS: P520 Astronoomia, kosmoseuuringud, kosmosekeemia

Using machine learning to improve physics-based photometric redshift estimations

Abstract: This thesis explores the application of machine learning methods to improve the accuracy of photometric redshift (photo-z) estimates by refining the outputs of the physics-based model TOPz. Using data from the WAVES survey, several regression-based machine learning models were trained to reproduce and enhance TOPz outputs. The best-performing model was XGBoost, which with an optimal configuration significantly reduced prediction errors and the proportion of outliers compared to the original TOPz estimates. Two enhancement models were developed: one directly predicted the logarithmic redshift transformation $\zeta = \ln(1 + z)$, while the other estimated the full ζ probability distribution. Both models were combined with the original TOPz outputs, with the direct ζ prediction model achieving the lowest error (MAE = 0,0265), and the probability distribution model providing better interpretability and the ability to handle ambiguous solutions. The best results were achieved by linearly or geometrically combining model outputs, optimizing the weight between TOPz and XGBoost contributions. The study demonstrates that a hybrid approach combining physical modeling and machine learning enables significantly more accurate and robust photometric redshift estimates, which are essential for large-scale astronomical surveys such as J-PAS and WAVES.

Keywords:
TOPz, photometric redshift, WAVES dataset, XGBoost, machine learning

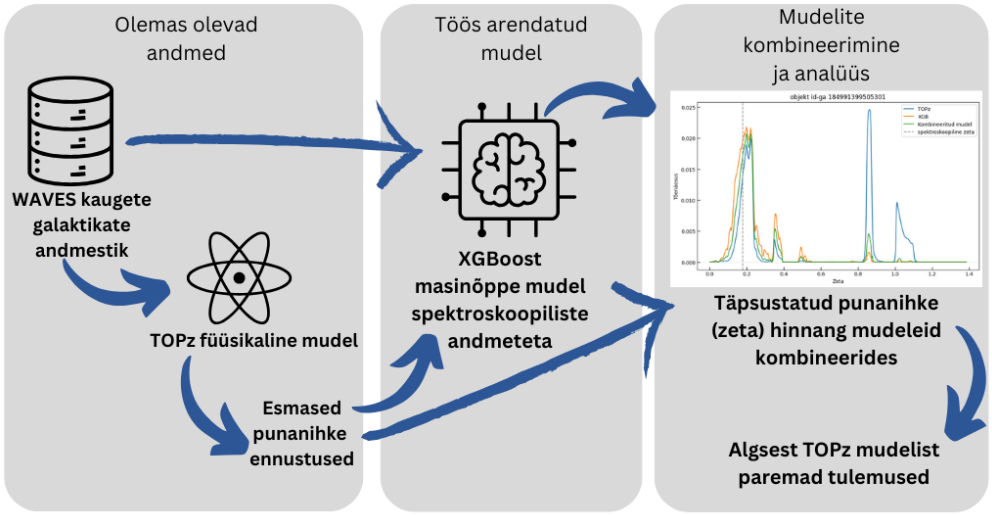
CERCS: P520 Astronomy, space research, cosmic chemistry

Masinõppe algoritmide rakendamine füüsikast lähtuvate fotomeetriliste punanihete täpsustamisel

TARTU ÜLIKOOL
arvutiteaduse instituut

Andmeteadus (MSc), 2025

Autor: Hannes Kuslap
Juhendajad: Elmo Tempel, PhD
Taavi Tuvikene, PhD



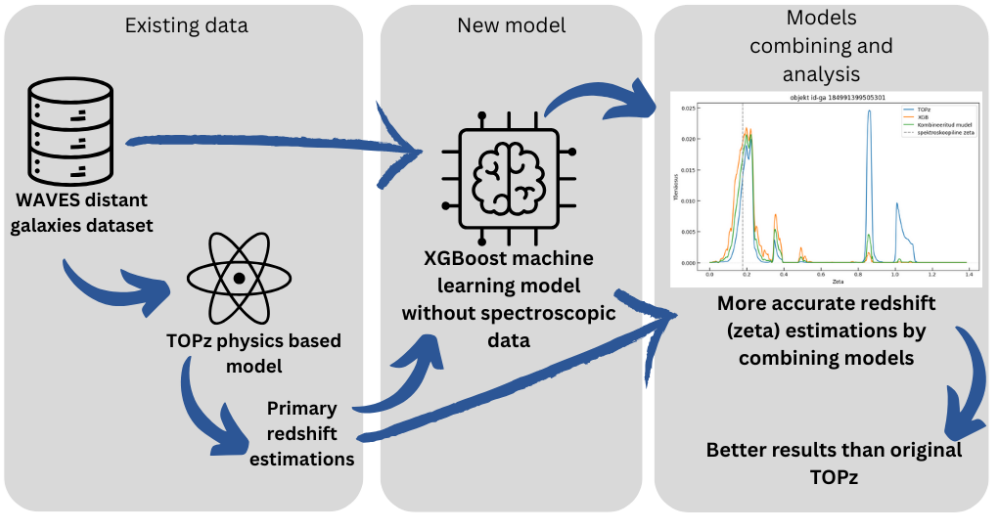
#UniTartuCS

Using machine learning to improve physics-based photometric redshift estimations

UNIVERSITY OF TARTU
Institute of Computer Science

Data Science (MSc), 2025

Author: Hannes Kuslap
Supervisors: Elmo Tempel, PhD
Taavi Tuvikene, PhD



#UniTartuCS

Sisukord

1 Sissejuhatus

Punanihe on objekti, eemaldumiskiirust kirjeldav suurus, mis näitab valguse lainepikkuse muutust sellest kiirusest tulenevalt. Kosmoloogiline punanihe, on astronoomilise objekti, näiteks galaktika nähtav punanihe, aga see ei tulene mitte enam objekti liikumisest, vaid ruumi paisumisest allika ja vaatleja vahel[Karttunen et al., 2007] . Kosmoloogilise punanihke põhjal saab ka hinnata kaugete objektide kaugust. Tavaliselt mõõdetakse punanihet spektrijoonte nihkumise järgi. Fotomeetriline punanihe on eemaldumiskiiruse hinnang, mis on saadud ilma objekti spektrit otseselt mõõtmata. See meetod kasutab fotomeetriat. Taevast pildistatakse erinevate filtritega, mis lasevad valgust läbi kindlates lainepikkuste vahemikes. Pilte analüüsides saadakse igale filtrile vastav galaktika heledus. Analüüsides, kuidas objektilt tulev valgus nendes vahemikes jaotub, saab hinnata objekti punanihet, võrreldes objekti mõõdetud heledusi eri filtrites teoreetiliste või empiiriliste mudelitega, mis kirjeldavad galaktikate värvide muutust punanihkega. Selline lähenemisviis võimaldab kiiresti ja efektiivselt hinnata suure hulga objektide punanihet, mistõttu on see meetod hädavajalik suuremahuliste taevauuringute ja universumi struktuuri ning arengu uuringute jaoks. Spektraalandmetega võrreldes on fotomeetrilisi andmeid palju suurema arvu galaktikate kohta, kuna nende mõõtmiseks kulub oluliselt vähem aega. Töös analüüsitav WAVES andmestik sisaldab nii spektroskoopilisi kui fotomeetrilisi andmeid.

Selles töös keskendun fotomeetrilistele andmetele. Töö eesmärk on täpsustada WAVES kataloogi andmetele TOPz algoritmiga leitud punanihke hinnanguid, kasutades selleks masinõppe mudeleid [Tempel et al., 2025]. TOPz (*Tartu Observatory Photo-z*) on TÜ Tartu observatooriumis välja töötatud fotomeetrilise punanihke (*photo-z*) hindamise tööriist, mis põhineb mallide sobitamisel ja Bayesi statistikal [Tempel et al., 2025]. Erinevalt masinõppe meetoditest, mis vajavad ulatuslikke spektroskoopilisi treenin-

gandmeid, kasutab TOPz füüsikalisi spektraalmalle, mis on genereeritud CIGALE tarkvaraga [Boquien, M. et al., 2019, Tempel et al., 2025]. See võimaldab hinnata galaktikate punanihkeid, võrreldes täheldatud fotomeetrilisi andmeid sünteetiliste spektritega [Tempel et al., 2025].

Varasemalt on ka otse masinõppega fotomeetrilisi punanihkeid leitud, näiteks kasutades ANNZ mudelit [Pathi et al., 2024]. Sellel meetodil esinevad aga puudused, kui ennustatav objekt väljub treeningandmete ruumist. Selle probleemi lahendamiseks on viimastel aastatel kombineeritud füüsikalisi mudeleid masinõppega ja leitud punanihkeid kombineeritud mudelite põhjal. Töös eesmärk on täpsustada TOPz algoritmiga saadud fotomeetrilisi punanihkeid, treenides nende peal masinõppe mudeli, mis suudaks eemaldada mõõtmisvigadest tekkinud erindeid.

Käesoleva töö teksti vormistamiseks ja silumiseks on kasutatud OpenAI chatGPT 4o ja 4.5 mudeleid.

2 Füüsikaline taust

2.1 Punanihe

Järgnev lõik tugineb Kartuneni õpikule [Karttunen et al., 2007]. Punanihe (tähistatud sümboliga z) on fundamentaalne mõiste astronoomias ja kosmoloogias, mis kirjeldab elektromagnetilise kiirguse lainepikkuse pikenemist, kui valgus liigub läbi laieneva universumi. See nähtus avaldub spektrijoonte nihkumisena pikemate lainepikkuste ehk punasema valguse suunas. Punanihe määratakse järgneva valemi abil:

$$z = \frac{\lambda_{\text{vaadeldav}} - \lambda_{\text{emiteeritud}}}{\lambda_{\text{emiteeritud}}} \tag{1}$$

kus $\lambda_{\text{vaadeldav}}$ on vaadeldud lainepikkus ja $\lambda_{\text{emiteeritud}}$ on allika poolt emiteeritud lainepikkus. Punanihe hõlmab endas erinevatest põhjustest tingitud lainepikkuse muutuseid, hõlmates Doppleri punanihet, mis tuleneb valgusallika ja vaatleja suhtelisest liikumisest, gravitatsioonilist punanihet, mis on tingitud valguse põgenemisest tugeva gravitatsioonivälja piirkonnast ning kosmoloogilist punanihet, mis on põhjustatud universumi laienemisest. Viimane on eriti oluline suurte kosmiliste kauguste ja varase universumi uurimisel. Kosmoloogiline punanihe tuleneb universumi enda laienemisest, mitte üksnes valgusallika ja vaatleja suhtelisest liikumisest. Valgus, mis pärineb kaugetelt galaktikatelt, venib universumi paisumise tõttu, põhjustades spektrijoonte nihkumise punasema spektri suunas. See nähtus peegeldab ruumi laienemist valguse teekonnal. Selle abil saab määrata kaugete objektide kaugusi ja nende taandumiskiirusi. Edwin Hubble avastas, et galaktikate punanihked on võrdelised nende kaugustega, mis viitas universumi laienemisele – see seos on tuntud kui Hubble'i seadus [Hubble, 1929]. Lisaks on punanihke mõõtmine oluline tööriist universumi vanuse, struktuuri ja arengu uurimisel.

Järgnev lõik tugineb Kartuneni õpikule [Karttunen et al., 2007]. Spektroskoopiline ja fotomeetriline punanihke mõõtmine on kaks peamist meetodit, mida kasutatakse galaktikate ja teiste taevakehade kauguste määramiseks, tuginedes valguse spektraalsete omaduste muutustele. Spektroskoopiline punanihe (z_{spec}) määratakse, mõõtes täpselt spektrijoonte nihkumist võrreldes nende laboratoorsete lainepikkustega. See meetod võimaldab väga täpseid mõõtmisi, kuna see tuvastab konkreetseid spektrijooni, mille nihkumine annab otsese teabe objekti liikumise ja kauguse kohta. Spektroskoopiline meetod on fotomeetrilisest usaldusväärsem, kuid nõuab palju teleskoobiaega ja on seetõttu piiratud heledamate objektide uurimisele. Fotomeetriline punanihe (z_{phot}) on hinnan-

guline meetod, mis põhineb objekti heledusmõõtmistel, kasutades laia läbilaskeribaga filtreid. See meetod võimaldab kiiresti määrata paljude galaktikate punanihkeid, eriti kui spektraalsete andmete kogumine ei ole praktiline. Kuigi fotomeetriline meetod on vähem täpne, on see kasulik suurte taevakehapopulatsioonide statistilisel analüüsil ja esialgsete kaugushinnangute saamisel.

Järgnev lõik tugineb Kartuneni õpikule [Karttunen et al., 2007]. Kokkuvõttes pakub spektroskoopiline punanihe suuremat täpsust ja on eelistatud meetod, kui on vaja täpseid kaugusmõõtmisi. Fotomeetriline punanihe võimaldab aga kiiret ja laiaulatuslikku objektide analüüsi, olles eriti kasulik suurte taevakehapopulatsioonide uurimisel, kus spektraalsete andmete kogumine oleks liiga ajamahukas või tehniliselt keeruline.

2.2 ζ kasutamine punanihke asemel

Järgnev lõik tugineb Baldry artiklile [Baldry, 2018]. Kosmoloogilistes analüüsides eelistatakse traditsioonilisele punanihkele z sageli logaritmilist lainepikkuse nihet, tähistatud kui ζ (zeta), mis on defineeritud järgnevalt.

$$\zeta = \ln(1 + z) \tag{2}$$

See eelistus tuleneb mitmest praktilisest ja teoreetilisest eelisest. Esiteks võimaldab ζ loomulikumat raamistikku kosmoloogiliste ja eriliikumise (ingl k. *peculiar velocity*) komponentide kombineerimiseks. Kuna punanihke mõõtmised esindavad nihkeid logaritmilisel lainepikkuse skaalal, võimaldab ζ kasutamine liita kiiruskomponente aditiivselt, lihtsustades erinevatel kosmilistel kaugustel olevate objektide liikumiste tõlgendamist [Baldry, 2018]. Teiseks pakub ζ fotomeetrilise punanihke hindamisel paremat numbrilist stabiilsust ja täpsust. Fotomeetrilised punanihke tehnikad, mis tuginevad laia

spektriga fotomeetriale, saavad kasu ζ logaritmilisest skaleerimisest, mis viib sümmeetrilisemate ja vähem kallutatud jääkideni fotomeetriliste ja spektroskoopiliste punanihete võrdlemisel. See on ka peamine põhjus, mis antud töös kasutan ζ väärtust punanihke asemel.

Järgnev lõik tugineb Baldry jt artiklile [Baldry, 2018]. Lisaks, kui analüüsitakse galaktikate ja suurte struktuuride jaotust, annab andmete kujutamine ζ funktsioonina ühtlasema jaotuse, eriti suurte punanihete korral. See ühtlus parandab galaktikate evolutsiooni ja klasterduse visualiseerimist ja tõlgendamist kosmilise aja jooksul. Samuti sobitub ζ kasutamine hästi universumi laienemise teoreetiliste mudelitega. Teatud kosmoloogilistes mudelites on kaasaliikuv kaugus, mis on võtmekontseptsioon universumi geomeetria ja laienemise mõistmisel, otseselt proportsionaalne ζ-ga, hõlbustades arvutusi ja tõlgendusi. Kokkuvõttes parandab teisenduse valemist 2 kasutamine kosmoloogilistes uuringutes punanihke mõõtmiste, kiiruse tõlgenduste ja universumi suurte struktuuride analüüsi selgust, täpsust ja tõhusust.

2.3 Fotomeetria

Kogu järgnev lõik on vastavalt Kartuneni õpikule [Karttunen et al., 2007]Astronoomias on fotomeetria meetod, mille abil mõõdetakse taevakehade, nagu tähtede, galaktikate ja planeetide, kiiratava valguse intensiivsust erinevates lainepikkuste vahemikes. See protsess hõlmab valguse kogumist teleskoobi abil ja selle suunamist läbi spetsiaalsete optiliste filtrite, et eraldada kindlad lainepikkuste vahemikud. Filtreeritud valgus registreeritakse seejärel seadmetega, nagu CCD-kaamerad või fotoelektrilised fotomeetrid, võimaldades kvantifitseerida taevakehade heledust. Fotomeetria on oluline erinevate astrofüüsikaliste omaduste, sealhulgas heleduse, temperatuuri ja koostise määramisel.

Kogu järgnev lõik on vastavalt Kartuneni õpikule [Karttunen et al., 2007]. Fotomeet-

riliste vaatluste standardiseerimiseks kasutatakse hästi määratletud filtrikomplekte, mida nimetatakse fotomeetrilisteks süsteemideks. Üks laialdasemalt kasutatavaid süsteeme on Johnson-Cousinsi UBVRI süsteem, mis sisaldab viit laia lainepikkusega filtrit: U (ultraviolett), B (sinine), V (visuaalne), R (punane) ja I (infrapuna). Need filtrid katavad optilise spektri ja on olulised tähtede klassifitseerimisel ning nende omaduste uurimisel, kuna on üks esimesi kasutusele võetud ja levinumaid fotomeetriliste filtrite süsteeme [Landolt, 2007]. Teine levinud süsteem on *Sloan Digital Sky Survey* (SDSS) *ugriz* süsteem, mis koosneb filtritest u, g, r, i ja z, ulatudes lähi-ultraviolettist lähi-infrapunani. SDSS filtrid on loodud pakkuma täpseid fotomeetrilisi mõõtmisi suurte taevavaatluste jaoks. Nende standardiseeritud filtrisüsteemide kasutamine võimaldab võrrelda vaatlusi erinevate instrumentide ja uuringute vahel, hõlbustades ühtset arusaamist universumist.

Kogu järgnev lõik on vastavalt Kartuneni õpikule [Karttunen et al., 2007]. Värvusindeks on fundamentaalne fotomeetriline parameeter, mis määratleb objekti heleduse erinevuse kahe erineva lainepikkuse filtris, näiteks $\mathrm{B} - \mathrm{V}$. See näitaja peegeldab objekti spektraalset energiajaotust (SED) ning sõltub nii objekti sisemistest omadustest, nagu temperatuur ja keemiline koostis, kui ka välistest teguritest, sealhulgas punanihkest. Kosmilise paisumise tõttu nihkub kaugemate galaktikate valgus pikematele lainepikkustele, muutes nende täheldatud värve. Seetõttu saab värvusindeksit kasutada punanihke kaudse mõõdikuna, kuna see kajastab punanihkest tulenevaid SED süsteemseid muutusi.

Kogu järgnev lõik on vastavalt Momtazi artiklile [Momtaz et al., 2022].Masinõppe kontekstis on värvusindeksid väärtuslikud tunnused fotomeetrilise punanihke hindamiseks. Need koondavad olulise spektraalse teabe kompaktseks kujuks, võimaldades masinõppemudelitel paremini tuvastada mustreid täheldatud värvide ja punanihke väärtuste vahel. Uuringud on näidanud, et värvusindeksite kaasamine masinõppe algoritmidesse parandab punanihke ennustuste täpsust. Näiteks SDSS andmete kasutamisel on leitud, et värvusindeksitele tuginevad mudelid saavutavad madalama standardhälbe punanihke

hindamisel, mis viitab suuremale täpsusele. Lisaks aitavad värvusindeksid vähendada fotomeetriliste vigade ja süsteemsete kallutuste mõju, kuna need on vähem tundlikud absoluutse kalibreerimise ebatäpsustele võrreldes toorheledustega. See muudab need eriti sobivaks suurte taevavaatluste jaoks, kus kõigi objektide spektraalsete punanihete määramine on ebapraktiline.

3 Metoodika

3.1 Punanihke hindamise meetodid

Järgnev lõik tugineb Merz jt artiklile [Merz et al., 2025]. Fotomeetrilised punanihke mõõtmise meetodid on üliolulised tänapäeva ja tulevaste suuremahuliste galaktikauuringute juures, sest pildiuuringute abil avastatakse galaktikaid palju kiiremini, kui on võimalik nende spektreid mõõta. Aastakümnete jooksul on välja töötatud hulgaliselt meetodeid fotomeetrilise punanihke arvutamiseks, mis üldiselt jagunevad kaheks peamiseks kategooriaks: mallipõhised meetodid (*template fitting*) ja masinõppepõhised meetodid.

Mallipõhise fotomeetrilise punanihke määramise korral võrreldakse vaadeldud objekti mitme lainepikkuse (filtri) fotomeetrilisi heledusi teadaolevate SED mallidega. Meetod proovib leida, millise spektrimalli ja punanihke kombinatsioon vastab kõige paremini vaadeldud fotomeetrilistele punktidele. Mallid võivad olla nii empiirilised (teiste galaktikate tegelikud spektrid) kui ka sünteetilised (teoreetilised galaktikamudelid) [Brescia et al., 2021] . Iga malli spekter nihutatakse erinevatele z väärtustele ning arvutatakse sellest tulenevad oodatavad heledused filtrites; neid võrreldakse vaatlustega ning hinnatakse sobivus (nt χ^2 minimaalsuse või maksimaalse tõenäosuse kriteeriumiga) [Brescia et al., 2021] . Parima sobivusega mall ja punanihe annavadki fotomeetrilise punanihke hinnangu. Sageli rakendatakse ka Bayesi lähenemist, lisades eelinfo (*prior*) galaktikate jaotuse kohta (nt heledusjaotus kui funktsioon punanihkest), et eelistada füüsikaliselt tõenäolisemaid lahendusi [Brescia et al., 2021]. Mallipõhise meetodi tulemusena saadakse lisaks punanihkele tihti ka objekti hinnanguline spektraalklass (millise malliga objekt seostus) ning punanihke tõenäosusjaotus (PDF), mis on väärtuslik info tulemuse statistilisest usaldusväärsusest [Brescia et al., 2021]. Seda tüüpi mudel on ka töös baasmudelina kasutuses olev TOPz mudel, mis arendati välja J-PAS vaatlusandmete

analüüsiks [Laur, J. et al., 2022].

Masinõppepõhised fotomeetrilise punanihk hindamise meetodid käsitlevad punanihke määramist regressioonülesandena, kus sisendiks on objekti fotomeetrilised parameetrid (enamasti heledused või värvusindeksid mitmes filtris) ning väljundiks punanihe. Mudel õpib seose sisendi ja väljundi vahel etteantud treeningandmete põhjal – selleks on tarvis suurt hulka objekte, millel on nii mitmefiltriline fotomeetria kui ka teada olev spektroskoopiline punanihe [Momtaz et al., 2022]. Õpitud mudelit saab seejärel rakendada uutele objektidele, et ennustada nende punanihet puhtalt fotomeetria põhjal. Masinõpe võib olla juhendatud, nagu eelkirjeldatud, aga on uuritud ka juhendamata meetodeid punanihete hindamise abistamiseks [Masters et al., 2015], [Hildebrandt, H. et al., 2010]. Levinumad algoritmid on näiteks tehisnärvivõrgud, kNN meetodid, toevektor-masinad (SVM), otsustuspuud ja juhuslikud metsad, Gaussi protsessid jpt [Momtaz et al., 2022]. Erinevad algoritmid pakuvad erinevaid eeliseid: näiteks närvivõrgud ja puud võivad automaatselt leida mittelineaarseid seoseid paljude tunnuste vahel, samas kui kNN või SVM võivad olla lihtsamini interpreteeritavad väiksema tunnuste arvuga andmestikel. Masinõppe rakendamine fotomeetrilise punanihke hindamise probleemile algas laiemalt 2000ndate alguses; üks esimesi populaarseid vahendeid oli ANNz, mis kasutas tehisnärvivõrku galaktikate punanihete prognoosimiseks [Pathi et al., 2024]. Selle uuendatud versioon ANNz2 lisas ka punanihke PDF-i hinnangu võimaluse ning on olnud kasutusel mitmetes projektides [Sadeh et al., 2016].

Mõlemal meetodil on omad eelised ja puudused. Masinõppe meetodite peamiseks eeliseks on nende empiriline täpsus, kui kasutusel on küllaldane ja esinduslik treeningandmestik. Arvukad uuringud näitavad, et kui treeningandmete hulk katab hästi fotomeetrilise parameetrite ruumi ja punanihete vahemiku, võivad masinõppe mudelid ennustada punanihkeid täpsemalt kui mallipõhised meetodid [Brescia, M. et al., 2014, Cavuoti, S. et al., 2012, Hildebrandt, H. et al., 2010]. Mallipõhise lähenemise suurim

pluss on selle füüsikaline üldistusvõime. Kuna meetod ei õpi ühelt kindlalt treeningvalimilt, vaid toetub füüsikalistele spektrimallidele, saab seda rakendada ka objektidele, mille tüüpi või heledusvahemikku pole varem spektroskoopiliselt kaardistatud. Teisisõnu puudub mallimeetoditel põhimõtteliselt piir fotomeetrilises sügavuses – neid saab kasutada ka väga nõrkade objektide puhul, kust spektroskoopia pole kättesaadav, eeldusel et objekt on siiski tuvastatud ja fotomeetrilised mõõtmised on olemas [Calzetti, 2014].

Mallipõhised ja masinõppepõhised meetodid pakuvad fotomeetrilise punanihke hindamisel erinevaid tugevaid külgi ning nende puudused on teineteist teatud mõttes täiendavad [Brescia, M. et al., 2014, Cavuoti, S. et al., 2012, Hildebrandt, H. et al., 2010]. Mallimeetod on põhimõtteliselt rakendatav iga objekti puhul (ka väga nõrkadel ja kaugetel, mida treeningandmetes pole) ning põhineb füüsikalistel mudelitel, kuid kannatab mallide ebatäiuslikkuse käes. Masinõpe suudab andmetest õpitavat infot ära kasutada maksimaalse täpsuse saavutamiseks ja on efektiivne suurtel andmehulkadel, kuid ebaõnnestub lihtsasti juhul, kui objekt erineb treeningandmetest või kui treeningandmed on kallutatud. Seetõttu ongi modernne lähenemine sageli kombineerida mõlemat tüüpi meetodeid, et ühendada nende tugevused ja kompenseerida nõrkusi. Näiteks hübriidmudelites kasutatakse masinõppe algoritme algul fotomeetrilise punanihke ennustamiseks ja seejärel tulemuse korrigeerimiseks mallisobitusega (või vastupidi) [Fotopoulou, S. ja Paltani, S., 2018]. Teine näide on hierarhiline Bayesi kombineerimine: seda on kasutanud mitme eri meetodi nii masinõppe kui mallsobituse mudelite punanihkehinnangute ühendamiseks Bayesi raamistikus, saades hinnangu, mis on täpsem kui ükskõik milline algsetest eraldi [Hatfield et al., 2022]. Neid viimast kahte meetodit kasutan ja võrdlen ka TOPz mudeli väljundi täpsustamisel masinõppe meetoditega.

3.2 TOPz

Järgnev lõik tugineb Elmo templi jt artiklile [Tempel et al., 2025]. TOPz (*Tartu Observatory Photo-z*) on Tartu Observatooriumis välja töötatud fotomeetrilise punanihke hindamise tööriist, mis põhineb mallide sobitamisel ja Bayesi statistikal. Erinevalt masinõppe meetoditest, mis vajavad ulatuslikke spektroskoopilisi treeningandmeid, kasutab TOPz füüsikalisi spektraalmalle, mis on genereeritud CIGALE tarkvaraga. See võimaldab hinnata galaktikate punanihkeid, võrreldes täheldatud fotomeetrilisi andmeid sünteetiliste spektritega. Lisaks sisaldab TOPz süsteem fotomeetriliste voogude ja nende määramatuste korrigeerimist ning füüsikalisi eeltingimusi, mis põhinevad galaktikate heledusfunktsioonidel, et parandada punanihke hinnangute täpsust ja usaldusväärsust.

Järgnev lõik tugineb Elmo templi jt artiklile [Tempel et al., 2025]. TOPz töövoog koosneb mitmest etapist. Esmalt luuakse lai spektraalmallide komplekt, mis katab erinevaid galaktikatüüpe ja arenguetappe. Seejärel optimeeritakse see komplekt, valides kõige esinduslikumad mallid, et parandada sobitamise täpsust. Täheldatud fotomeetrilised vood ja nende määramatused korrigeeritakse süsteemsete vigade vähendamiseks. Lisaks rakendatakse analüütilist eeltingimust, mis põhineb galaktikate jaotusel universumis, et täpsustada punanihke tõenäosusjaotusi. Lõpuks annab TOPz mitte ainult punanihke hinnangud, vaid ka nendega seotud määramatused ning lisaks galaktikate tähtmasside hinnangud.

TOPz-i rakendamine GAMA (Galaxy And Mass Assembly) uuringu andmetele näitas selle meetodi kõrget täpsust[Tempel et al., 2025]. Võrreldes spektroskoopiliste punanihetega, olid TOPz-i fotomeetrilised hinnangud minimaalse nihkega ja väikese hajuvusega, andes samaväärseid või paremaid tulemusi teiste tuntud fotomeetrilite punanihke hindamise algoritmidega nagu EAZY või SFM [Tempel et al., 2025]. Lisaks näitasid TOPz-i poolt hinnatud massid tugevat kooskõla spektroskoopiliste meetoditega saadud

väärtustega[Tempel et al., 2025]. Need tulemused rõhutavad TOPz-i võimekust pakkuda täpseid ja usaldusväärseid punanihke ja massi hinnanguid, muutes selle väärtuslikuks tööriistaks praegustes ja tulevastes fotomeetrilistes uuringutes, sealhulgas planeeritud rakendustes J-PAS kitsa lainepikkusega uuringus [Tempel et al., 2025, Laur, J. et al., 2022].

3.3 Andmete kirjeldus

Töös kasutan WAVES kataloogi andmeid, mis on eelnevalt puhastatud TOPz mudeli jaoks, seega saan eeldada, et kõik andmepunktid on olemas ja realistlikus suurusjärgus. Antud kataloogis on üle 25 miljoni objekti, aga töös kasutan vähendatud ja puhastatud andmestikku, kuhu on jäetud vaid 295 684 objekti, et oleks analüüsimiseks ka spektroskoopilised punanihked alles [Driver et al., 2016]. TOPz mudel annab väljundina mitu faili, millest valin enda mudeli jaoks sobilikuimad ja vajalikud. Oma mudeli jaoks kasutan lõpuks 3 faili, mille sisu järgnevalt kirjeldan. Esmalt loen sisse zeta väärtuste tõenäosusjaotuse failist *topz_hannes_cat_pzeta.fits*. See fail sisaldab 5 tulpa: *obj_id, obj_name*, *zspec*, *mag* ja *pzeta*. Tulbad *obj_id* ja *obj_name* on identsed ja nende abil saab tuvastada üksikobjekte ning neid siduda tabelite vahel. Tulp *zspec* on spektraalne punanihe, mille teisendan ise ümber suuruseks ζ_{spec} vastavalt füüsikalise tausta peatükis toodud valemile. Tulp mag näitab objekti tähesuurust ja ja seda pole siin töös kasutatud. Tulp pzeta on 500-kohaline vektor, mis näitab TOPz põhjal saadud tõenäosusjaotust, kus iga indeks vastab mingile zeta väärtusele. Selle teisenduse tegemiseks kasutan faili *topz_hannes_zeta.fits*, mis seab iga indeksi vastavausse sobiva ζ väärtusega. Lisaks annab see fail ka vastavale indeksile (ehk ka ζ väärtusele) objekti võimalikud distantsid, vanuse ja ruumala, kuid need pole antud töö seisukohast olulised. Viimaks loen sisse täiendavad treeningandmed failist *topz_hannes_catalog_Fmag.fits*. Sellest failist kasutan heledusi erinevates filtrites ja nende vigu. Selles andmestikus on filtreid 9.

3.4 Töövahendid

Antud töö teostasin Pythoni keeles Jupyter Notebook formaadis Google Colab keskonnas. Lisaks kasutasin Google Colabi sisseehitatud tehisintellektimudelit Google Copilot ja ChatGPT mudeleid GPT - 4o ja GPT - 4,5. Samuti kasutasin TOPCAT programmi fits-laiendiga failide uurimisel enne mudelte treenimist. Kasutasin ka järgnevaid vabavaralisi pythoni teeke: astropy, pandas, numpy, matplotlib, seaborn, time, sklearn ja xgboost.

3.5 Töö ülesanded

Töö eesmärk on treenida masinõppe mudel, mis treenib ennast TOPz väljundil ja parandaks TOPz täpsust. Varasemalt on mulle antud TOPz jaoks puhastatud andmestik, mida kirjeldasin eelnevalt andmete kirjelduse peatükis. Minu ülesanne on lugeda sisse andmed ja leida sobiv mudel, mis suudaks reprodutseerida TOPz väljundit täpselt ja arvutuslikult küllalt efektiivselt. Seejärel kontrollin kas kehtib hüpotees, et TOPz väljundit saab täpsemaks muuta, treenides mudeli treeningandmetena TOPz mudeli väljundina saadud ζ tõenäosusjaotuse maksimaalse väärtuse asukoha väärtust kasutades. Valideerin oma ennustusi spektroskoopilise punanihke väärtuste järgi.

Viimase sammuna tuleb mudelit muuta nii, et see treenitaks ennustama diskreetse ζ väärtuse asemel ζ tõenäosusjaotust. Mudeli väljundit tuleb siluda ja skaleerida sobilikuks. Selle tulemusena saadud uuele mudelile tuleb leida sobiv kombineerimise meetod TOPz mudeliga; selleks proovisin erinevaid tõenäosusjaotuste kombineerimise meetodeid. Lõpuks tuleb saadud kombineeritud mudeli headust valideerida ja kirjeldada spektroskoopiliste ζ_{spec} väärtuste vastu.

4 Mudeli kirjeldus

4.1 Andmete eeltöötlus

Töös kasutan andmeid, millele on juba TOPz algoritmi rakendatud, seega on enamik vajalikku eeltöötlust tehtud TOPz algoritmi rakendamise eelselt. Kuna kogu töö mõte on rakendada arendatud mudelit TOPz väljundile, siis ei ole vajalik andmete puhastamine, kuna võib eeldada, et nii need kui tulevased andmed on juba puhtad. Kuna andmed sisaldavad objektide tähesuuruseid erinevates filtrites, aga mitte värvusindekseid, lisan ka need. Kuna andmetes on heledus 9 erinevas filtris, siis saab arvutada 36 erinevat värvusindeksit.

4.2 Proovitud mudelid

Esimese sammuna uurin, milline mudel suudab kõige paremini reprodutseerida TOPz mudeli väljundit ehk TOPz väljundi tõenäosusjaotuse kõrgeima tõenäosusega ζ väärtust. Selle põhjal valin välja sobivaimad mudelid, mida kasutada edasi spektroskoopilise ζ_{spec} ennustamiseks.

Võrdlusesse võtsin esialgu 6 erinevat mudelit: *Linear Regression*, *Random Forest*, *Gradient Boosting*, *SVR*, *KNN Regressor* ja *XGBoost*. Kuna hiljem on vaja mudelit kasutada suuremal andmestikul, siis ajaressursi piiratuse tõttu jätan neist võrdlusesse alles vaid kiireimad mudelid: *Linear Regression*, *Random Forest*, *Gradient Boosting* ja *XGBoost*. Parimaid tulemusi andsid *XGBoost* ja *Random Forest*, aga kuna *XGBoost* on oluliselt kiirem, siis otsustasin edasi minna selle mudeliga ja otsisin sellel ka parimaid hüperparameetreid. Kuna *Random Forest* algoritm on teistest tunduvalt aeglasem, aga andis siiski häid tulemusi, proovisin seda ka teiste hüperparameetritega, kus puude arv

on 5, et oleks lähemal *XGBoosti* treeningajale. Selle tulemusena on näha, et endiselt oluliselt aeglasemalt andis see halvemaid tulemusi. Proovitud parameetrite ruum on järgnev: *n_estimators*: [50, 100, 200], *max_depth*: [3, 5, 7], *learning_rate*: [0,01; 0,1; 0,2], *colsample_bytree*: [0,3; 0,5; 0,7]. Parimateks parameetriteks osutusid *n_estimators* = 200, *max_depth* = 7, *learning_rate* = 0,2, *colsample_bytree* = 0,7. See mudel on tabelis 1 toodud kui XGBoost2. Samuti selgub mudeleid võrreldes, et enim mõjutab parameeter *n_estimators* Selles tabelis on esitatud kuue erineva masinõppemudeli võrdlus, mida on kasutatud TOPz mudeli reprodutseerimisel. Näitajateks on R^2 skoor (selgitusjõu mõõdik), MAE (keskmine absoluutne viga), MSE (keskmine ruutviga) ja treeninguaeg sekundites. Need tulemused võimaldavad hinnata, millised mudelid pakuvad parimat tasakaalu täpsuse ja arvutusliku efektiivsuse vahel. Siin ja edaspidi kasutan mudelite nimetusi konkreetsete treenitud mudelite nimedena, mis on toodud ka tabelites.

Tabel 1. Erinevate masinõppemudelite võrdlus TOPz mudeli reprodutseerimisel

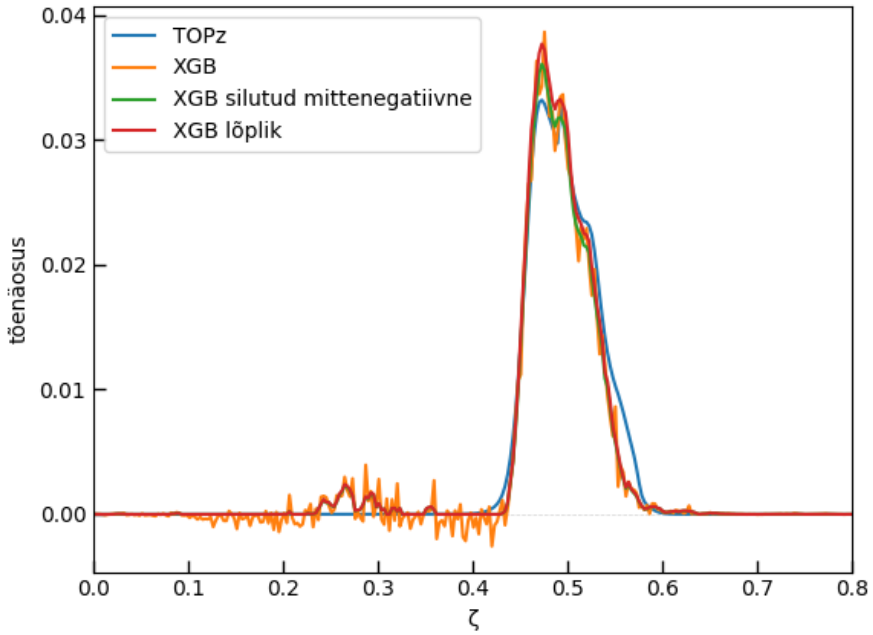
Mudel	**R^2 skoor**	**MAE**	**MSE**	**Treeningu aeg (s)**
Linear Regression	0,844	0,0361	0,00288	0,2
Random Forest	0,941	0,0163	0,00109	973,6
Random Forest2	0,926	0,0191	0,00136	47,8
Gradient Boosting	0,877	0,0319	0,00227	300,7
XGBoost	0,931	0,0206	0,00127	5,6
XGBoost2	0,937	0,0186	0,00115	5,9

4.3 Mudeli kasutamine

Kui tavaliselt jagatakse andmed mudeli treenimisel kaheks osaks, treening- ja testandmeteks, siis käesolevas töös on oluline mõista, et mudel treenib nähes ainult TOPz mudeli treeningandmeid ja TOPz väljundit ning selle põhjal täpsustab mudelit, seega toimub valimisisene ennustamine ja andmeid mitmesse osasse jagama ei pea. Kuna eelnevalt sai valitud parimaks mudeliks XGBoost, siis samuti jätkan järgnevalt selle mudeli ka-

sutamisega. Esmalt kasutan sama mudelit rakendades seda ennustama TOPz väljundi tõenäosusjaotusest saadud maksimumi asukoha ζ_{TOPz} väärtust. Kui eelmises osas tahtsin vaid võimalikult hästi reprodutseerida TOPz väljundit, et valida sobiv mudel, siis nüüd treenin mudeli kõigil andmetel ja väljundit kontrollin enam mitte ζ_{TOPz} vastu, vaid kombineerin lineaarselt ζ_{TOPz} väärtustega ning leian sobivad kaalud. Lõplikke saadud ζ väärtuseid võrdlen spektroskoopiliste ζ_{spec} väärtuste vastu. Võrdlesin täpsust ainult TOPz väljunditäpsusega ilma minu lisatud mudelita.

Kui see andis positiivse tulemuse, liikusin edasi tõenäosusjaotuse mudeli ehitamise juurde. Erinevus eelnevast mudelist on see, et siin mudel peab ennustama ζ asemel zeta tõenäosusjaotust. TOPz mudeli üks väljunditest on ka ζ_{TOPz} tõenäosusjaotus, kus on ζ väärtused diskretiseeritud 500 väärtuse vahel vahemikus 0 kuni 1,385. Et enamik sisendeid on samad eelneva ülesandega ja eelnevates ülesannetes toimis kõige efektiivsemalt XGBoost, siis kasutan ka siin XGBoosti mudelit, aga proovin ka erinevaid puude arve, et leida parim võimalik mudel sellele ülesandele. Kuna siin oli vaja leida tõenöosusjaotus ehk ennustada 500 erinevat väärtust, proovisin 2 varianti. Tavaline XGBoost n-mõõtmelise väljundi jaoks kasutab n erinevat mudelit. Teine võimalus on lisada parameeter multi strategy='multi output tree'. Proovisin mõlemaid variante.

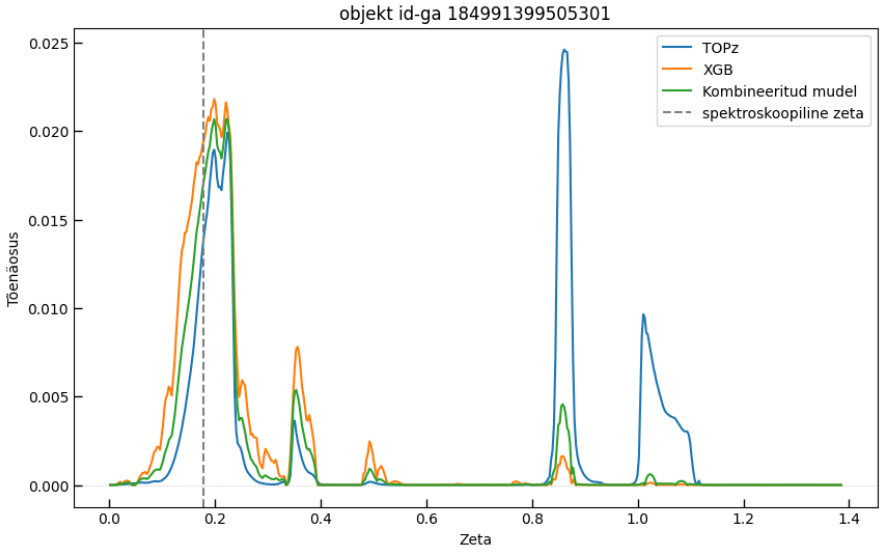


Joonis 1. Tõenäosusjaotusi ennustava XGBoosti ennustus ja selle töötlemine võrreldud TOPz ennustusega.

Oma mudeli ennustuse toorväljundil esinesid aga mõningad vead, näiteks sisaldas see negatiivseid väärtuseid. Ennustuse tõenäosusjaotusel tehtavad protsessid kujutasin ühe objekti puhul joonisel 1. Sinisega on seal kujutatud TOPz ennustus ja oranžiga XGBoost1000 ennustus. XGBoost1000 mudelile vastavalt oranžil joonel on näha suurt müra, mille eemaldamiseks kasutan silumist, sest ennustus ei tea, et tegemist on sõltuvate suurustega järjestikkuste punktide puhul. Kasutan silumist Savitsky-Golay meetodiga [Savitzky ja Golay, 1964]. Savitzky–Golay silumismeetod on levinud digitaalsete signaalide töötlemise tehnika, mida kasutatakse andmete müra vähendamiseks, säilitades samal ajal andmestruktuuri olulised omadused, nagu lokaalsed kõverused, tippude asukohad ja kuju. Rakendasin Savitzky–Golay silumist parameetritega akna pikkus $w = 7$ ja polünoomi aste $p = 2$, mis osutus sobivaks kompromissiks müra tasandamise ja signaali struktuuri säilitamise vahel. Valitud akna suurus tähendab, et iga punkti väärtus arvutatak-

se, sobitades selle ümber paiknevatele seitsmele punktile teist järku polünoomi. Selline konfiguratsioon võimaldab siluda lokaalseid kõikumisi ilma oluliselt mõjutamata signaali üldist kuju või väiksemaid lokaalseid omadusi. Eriti oluline on see just tõenäosusjaotuste korral, kus tipu kuju ja asukoht on olulised ennustuse täpsuse hindamisel. Silumismeetodi rakendamine võimaldas saavutada sujuvamaid ja statistiliselt usaldusväärsemaid jaotusi, mis omakorda parandas hilisemates etappides tippude tuvastamist ning mudelite võrdluse tulemusi.

Lisaks silumisele lisasin tingimuse, et kõik punktid peavad olema mittenegatiivsed, et vältida negatiivseid tõenäosuseid; seda tingimust kasutav tõenäosusjaotus on joonisel 1 kujutatud rohelise joonega. Lõpuks, et TOPz ja XGBoost mudelid oleks kombineerimisel konstantse kaaluga, skaleerisin XGBoosti ennustuse summale 1. Selleks jagasin iga punkti väärtuse läbi vektori ennustuste summaga. Selle tulemusena sain XGBoosti lõpliku ennustuse, mis on joonisel 1 kujutatud punase joonega.



Joonis 2. Tõenäosusjaotusi ennustava XGBoosti ennustus ja selle töötlemine võrreldud TOPz ennustusega.

Sarnaselt otse ζ väärtuse ennustamise mudeliga kombineerisin ka siin mudelit TOPz väljundiga, et tekitada kombineeritud mudel. Mudelite kombineerimiseks katsetasin lineaarset ja geomeetrilist keskmist ning leidsin parima meetodi ja parimad kaalud. Kombineerimise tulemus näidisobjekti puhul on kujutatud joonisel 2, Joonisel on näha olukord, kus üks mudel ennustab teisega võrreldes mingile ζ väärtusele oluliselt madalamat tõenäosuse väärtust ja seeläbi kombineeritud mudelis saab mõjule teine lokaalne maksimum. Kombineeritud mudelit ma enam ei skaleeri, kuna see ei muuda maksimumi asukohta ja seega ei muuda tulemusi. Tõenäosusjaotuste mudel on vajalik, et paremini kirjeldada, miks mingi ennustus tehakse, kuna sellel saab vahetult näha, kuidas erinevad mudelid ennustavad ζ tõenäosusjaotust, nagu näha jooniselt 2.

5 Tulemused ja arutelu

5.1 Tõenäoliseima ζ väärtusel mudel

Selles peatükis käsitlen mudelit, mis saab ette TOPz väljundi vaid tõenäoliseima ζ väärtusena, ehk TOPz väljundi tõenäosusjaotusest on valitud kõrgeima tõenäosusega ζ väärtus. Proovin samu mudeleid, mis eelnevalt olen välja valinud TOPz väljundi reprodutseerimisel, kuid kuna eelnevalt oli näha, et parimaid tulemusi annavad XGBoost mudelid, siis lisan neid juurde muutes hüperparameetreid. Kuna hüperparameetrite analüüsil selgus, et enim mõjutab mudeli väljundit parameeter n_estimators, siis ka siin osas proovin erinevaid selle suuruse väärtuseid, lisades katsesse väärtused 100, 200, 1000 ja 5000, Samuti jätan ka varasemad mudelid kasutusse alles.

Alljärgnevalt on esitatud analüüs tabelis 2 toodud masinõppemudelite jõudluse kohta, võrreldes neid algse TOPz mudeliga spektroskoopiliste andmete põhjal. Mudelite hindamisel kasutati peamiste kvaliteedimõõdikutena determinatsioonikordajat (R^2), keskmist absoluutviga (MAE), keskmist ruutviga (MSE), erindite osakaalu ja mudelite treeninguaega sekundites. Peamiseks eesmärgiks on valida mudel, mis tagaks optimaalse kombinatsiooni ennustustäpsusest ja arvutuslikust efektiivsusest, et seda saaks kasutada tulevikus rohkemate filtritega või rohkemate andmetega andmestikel. Antud töös on kasutuses 295 684 objekti enam kui 25 miljonist WAVES kataloogis olevast andmepunktist, millele plaanitakse mudelit tulevikus rakendada [Driver et al., 2016].

Mudelitest paistavad silma Random Forest ja XGBoost meetodil põhinevad lahendused, saavutades kõige paremad täpsusnäitajad. Kuigi Random Forest mudelil oli kõrgeim R^2 väärtus (0,916) ja väikseim MSE (0,00162), osutus see arvutuslikult äärmiselt ressursimahukaks (treeningaeg üle 1368 sekundi ehk ligi 23 minutit). Selline pikk treeningaeg muudab Random Foresti praktilistes rakendustes ebasobivaks, kuna andmestik, kus

Tabel 2. Erinevate masinõppe mudelite võrdlus TOPz mudeliga spektroskoopiliste andmetega võrdlusel

Mudel	**R^2 skoor**	**MAE**	**MSE**	**Erindite osakaal**	**Treeningaeg (s)**
TOPz	0,891	0,0275	0,00210	0,0215	-
Linear Regression	0,859	0,0372	0,00272	0,0305	0,2
Random Forest	0,916	0,0256	0,00162	0,0162	1368,4
Random Forest2	0,908	0,0267	0,00178	0,0180	69,6
Gradient Boosting	0,877	0,0356	0,00237	0,0246	473,6
XGBoost100	0,914	0,0274	0,00167	0,0140	6,1
XGBoost200	0,914	0,0265	0,00166	0,0140	11,5
XGBoost1000	0,907	0,0260	0,00181	0,0168	47,8
XGBoost5000	0,896	0,0268	0,00201	0,0204	227,8

mudelit kasutatakse, on kordades suurem antud töös analüüsitavast andmestikust.

XGBoost-põhised mudelid pakkusid oluliselt tasakaalustatumat lahendust. XGBoost200 mudel, mida treeniti 200 iteratsiooniga, saavutas väga kõrge determinatsioonikordaja väärtuse ($R^2 = 0,914$), mis oli madalam kui Random Forestil, kuid märgatavalt parem võrreldes baaslahendusega ehk TOPz mudeliga ($R^2 = 0,891$). Samuti olid XGBoost200 mudeli veanäitajad väga head (MAE = 0,0265 ja MSE = 0,00166), olles ühed madalaimad võrreldes kõigi teiste mudelitega. See tähendab, et XGBoost200 suutis teha täpseid ennustusi ja minimeerida ennustusvigade suurust.

Lineaarregressiooni ja Gradient Boostingu tulemuste analüüs näitas selgelt, et nende mudelite ennustusvõime jäi oodatust kehvemaks ning ei suutnud ületada isegi algse TOPz mudeli taset. Näiteks lineaarregressiooni determinatsioonikordaja ($R^2 = 0,859$) jäi oluliselt alla TOPz tulemusele ($R^2 = 0,891$), mis viitab mudeli piiratud võimele tabada andmestiku keerulisi mustreid ja mittelineaarseid seoseid. Ka Gradient Boostingu mudel, vaatamata üldiselt heale kohandumisvõimele erinevate andmekogumitega, näitas käesolevas töös mõõdukat jõudlust ($R^2 = 0,877$), jäädes samuti alla TOPz-le. Sellised tulemused näitavad, et nimetatud meetodid ei pruugi antud andmekogumi eripärasid

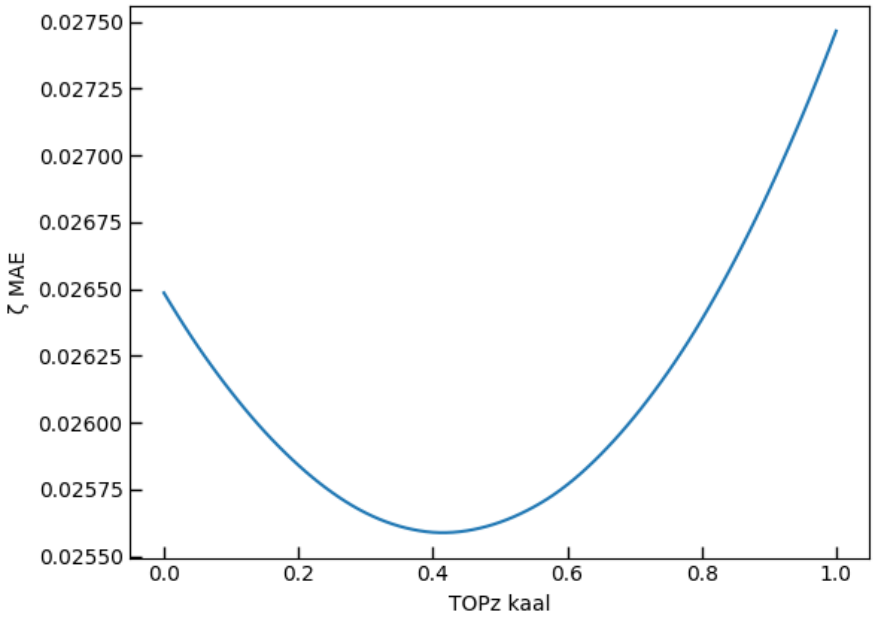
ja mittelineaarseid seoseid piisavalt hästi ära tabada ning nende praktiline kasutamine spektroskoopiliste andmete analüüsis ei ole soovitatav. Seevastu kõik ülejäänud testitud mudelid, sealhulgas Random Forest ja erinevad XGBoosti variandid, näitasid paremat jõudlust kui TOPz mudel, kinnitades mittelineaarsete masinõppemeetodite sobivust antud ülesandele.

Lisaks täpsusele oli XGBoost200 mudeli treeningaeg (11,5 sekundit) teiste kõrge jõudlusega mudelitega võrreldes märkimisväärselt lühem. Näiteks oli see oluliselt kiirem võrreldes Gradient Boosting (473,6 sekundit) ja Random Forest mudelitega. Kuigi leidus ka kiiremaid mudeleid, näiteks väiksema iteratsioonide arvuga XGBoost100 (6,1 sekundit), jäi selle mudeli jõudlusnäitajad mõnevõrra kehvemaks, mis näitab, et XGBoost200 saavutas optimaalse kompromissi arvutusliku kiiruse ja tulemuste täpsuse vahel.

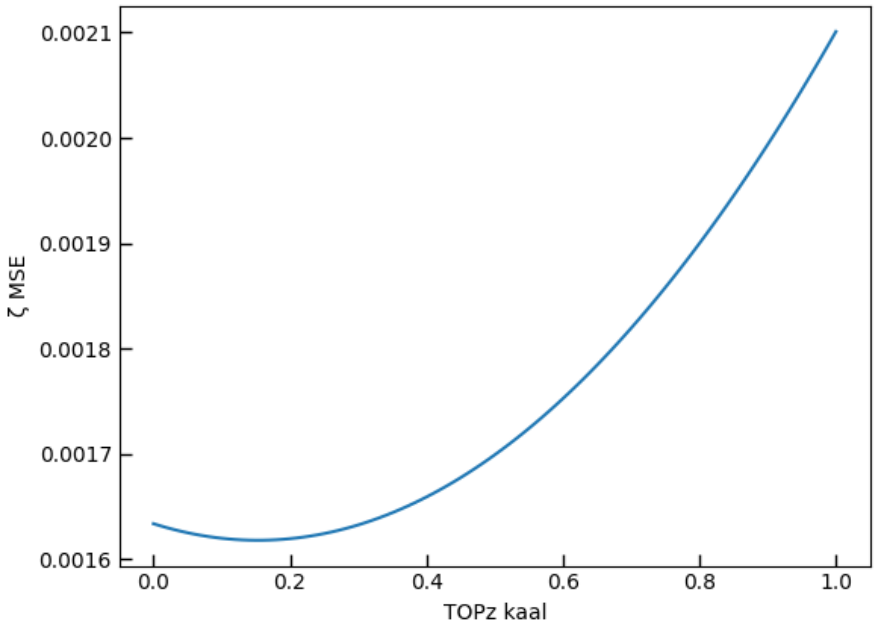
XGBoost200 ja Random Foresti mudeli otsene võrdlus toob esile olulised nüansid mudelite praktilise rakendatavuse seisukohast. Kuigi Random Forest saavutas marginaalselt parema determinatsioonikordaja ($R^2 = 0,916$ vs. $R^2 = 0,914$ XGBoost200-l), oli see arvutuslikult äärmiselt ressursimahukas, võttes treenimiseks üle 1300 sekundi. Selline pikk treeninguaeg piirab Random Foresti mudeli praktilist kasutatavust. Lisaks, kuna käesolevas töös on ka oluline minimeerida erindite osakaalu, tuleb rõhutada, et XGBoost200 mudeli erindite osakaal (0,0140) oli madalam kui Random Forestil (0,0162) ning märkimisväärselt parem kui TOPz mudelil (0,0215). Just erindite vähesus on oluline, kuna kõrge erindite osakaal muudab mudeli ebakindlaks ja vähendab selle praktilist usaldusväärsust. Seetõttu on XGBoost200 selgelt eelistatud lahendus, pakkudes head kompromissi täpsuse, madala erindite hulga ning arvutusliku efektiivsuse vahel.

Eelnevalt kirjeldatud põhjuste tõttu valiti käesolevas töös parimaks mudeliks XGBoost200. Antud mudel pakub suurepärase kombinatsiooni täpsusest ja arvutuslikust tõhususest, mis teeb selle sobivaks praktiliste rakenduste jaoks, kus vajalik on kõrge

ennustusvõimekus, samas hoides treeningaja ja ressursikasutuse mõistlikul tasemel. Mudeli optimeerimiseks tehtud iteratsioonide arv (200) oli piisav, et saavutada kõrge täpsus, vältides samal ajal liigset arvutusressursside raiskamist. Kokkuvõttes toetavad tulemused selgelt XGBoost200 mudeli valikut antud töö seisukohast. Seega teen järgnevad analüüsid kasutades mudelit XGBoost200, mis on XGBoost mudeli parameetritega (objective =’reg:squarederror’, colsample_bytree = 0,7, learning_rate = 0,2, max_depth = 7, n_estimators = 200).



Joonis 3. Kombineeritud mudeli MAE TOPz ja XGBoost1000 lineaarse kombinatsiooni TOPz kaalust sõltuvana.



Joonis 4. Kombineeritud mudeli MSE TOPz ja XGBoost1000 lineaarse kombinatsiooni TOPz kaalust sõltuvana.

Järgnevalt on esitatud mudelite XGBoost200 ja TOPz lineaarne kombineerimine, kus lõplik ennustusväärtus ζ leitakse kaalutud summana järgnevast valemist.

$$W_1 \cdot \zeta_{\mathrm{TOPz}} + (1 - W_1) \cdot \zeta_{\mathrm{XGBoost}} \tag{3}$$

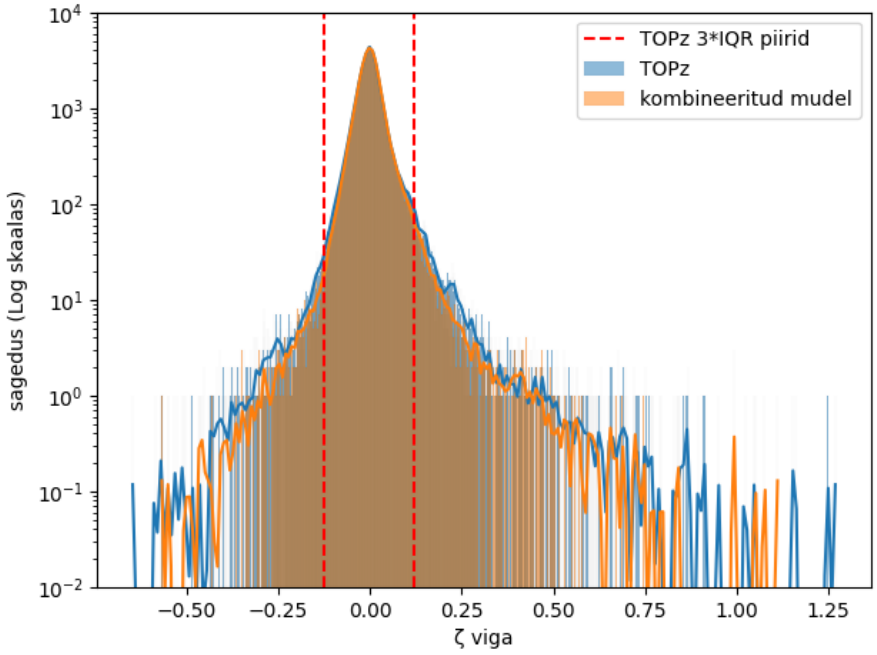
Antud lähenemise eesmärgiks on kasutada ära mõlema mudeli tugevusi ning vähendada veelgi lõppennustuse vigasid. Mudelite kombineerimise tulemusena saadud sõltuvusi kaalutegurist W_1 on visualiseeritud joonistel 3 ja 4, kus on kujutatud kombineeritud mudeli ennustuse ζ keskmise absoluutvea (MAE) ja keskmise ruutvea (MSE) muutused sõltuvalt kaalutegurist W_1.

Graafikute 3 ja 4 põhjal valiti optimaalse kaaluteguri W_1 määramisel peamiseks

hindamiskriteeriumiks keskmine absoluutviga, kuna see näitab kõige selgemalt mudelite praktilist ennustusvõimekust ning võimaldab andmete analüüsis tõhusalt hinnata, kui täpselt ennustatud väärtused vastavad tegelikele mõõtmistulemustele. MSE kasutamine oleks küll samuti sobilik, kuid MAE eelistamine tulenes selle robustsemast käitumisest üksikute suuremate vigade suhtes, mis on käesoleva töö eesmärkide seisukohalt oluline, sest need on TOPz puhul levinud.

Tabelist 2 on selgelt näha, et XGBoost200 mudel annab oluliselt vähem erindeid kui algne TOPz mudel. Kuna TOPz mudeli ennustustes esineb suuremaid üksikuid kõrvalekaldeid ehk erindeid, mõjutavad need eriti tugevalt just keskmist ruutviga, mis annab suurematele vigadele teise astme tõttu suure osakaalu. Seetõttu näeme graafikul 4 optimaalset TOPz kaalu madalamana võrreldes MAE-ga, sest suuremate erindite tõttu on kasulik anda XGBoost200 mudelile suurem osakaal lõpliku kombineeritud mudeli koostamisel. Niisiis tuleneb väiksem optimaalne TOPz kaal MSE joonisel otseselt XGBoost200 mudeli väiksemast erindite arvust ja stabiilsemast ennustusvõimekusest võrreldes TOPz-ga.

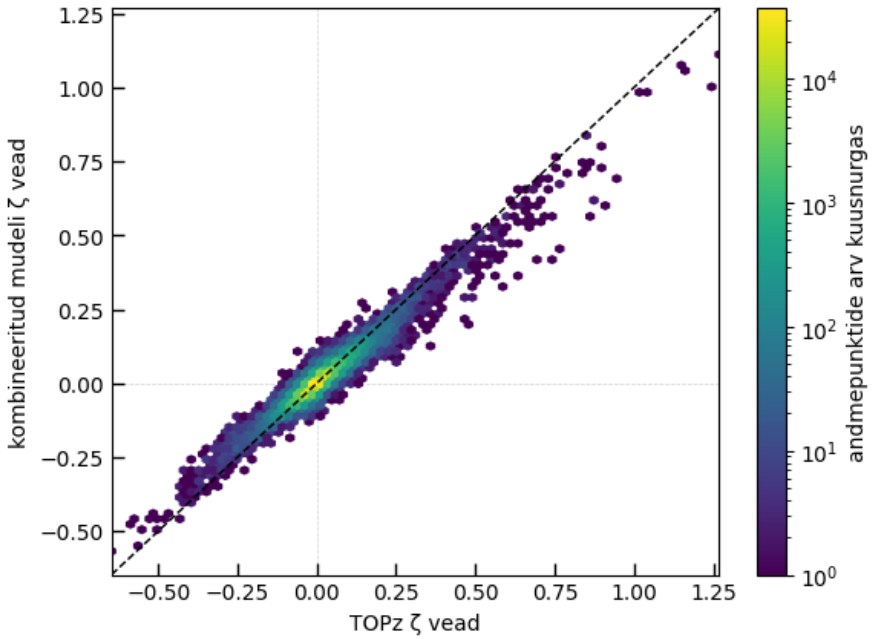
Tulemustest on selgelt näha, et eksisteerib optimaalne kaalutegur, mille korral on lõplik keskmine absoluutviga minimaalne. Graafiku põhjal selgub, et optimaalne kaalutegur W_1 paikneb vahemikus 0,35 – 0,45, mis tähendab, et lõplikus mudelis on mõistlik kombineerida mõlemaid meetodeid ligikaudu võrdsete kaaludega. Edasisteks arvutusteks kasutasin TOPz kaaluga 0,41, ja XGBoost200 vastavalt kaaluga 0,59, mis on valitud joonise 3 miinimumi asukoha järgi. Selline kombinatsioon võimaldab ära kasutada XGBoost200 mudeli suurt üldist täpsust ja TOPz mudeli spetsiifilisi omadusi, mis aitavad vähendada lõplikku ennustusviga ja suurendada mudeli praktilist usaldusväärsust spektroskoopiliste andmete analüüsil.



Joonis 5. TOPz ja kombineeritud mudeli ζ ennustamise vigade jaotus võrreldes spektroskoopilise ζ_{spec} väärtusega.

Joonisel 5 on kujutatud TOPz mudeli ja kombineeritud mudeli (XGBoost200 ja TOPz) vigade jaotus logaritmilisel skaalal. Veatelgedele on lisatud ka punase joonega tähistatud TOPz mudeli põhjal määratud $3 \cdot$ IQR (interkvartiilvahemiku) piirid, mille ületamist käsitletakse Tukey kriteeriumi põhjal erindina. Tabeli 2 põhjal teame, et erindeid on ligikaudu 1-2%, seega $3 \cdot$IQR joonte vahele jääb tegelikult ligikaudu 98% kogu andmetest. Jooniselt on näha, et kombineeritud mudeli (oranž) vigade jaotus on koondunum nulli ümber ning selle sabad on oluliselt madalamad võrreldes TOPz mudeliga (sinine joon), mis näitab väiksemat suurte vigade (erindite) arvu. Väga kitsas 0 ümbruse vahemikus on ka näha, et TOPz KDE joon on kõrgemal kui kombineeritud mudeli oma, mis näitab, et TOPz saab rohkematel juhtudel väga täpselt pihta ζ väärtusele kui kombineeritud mudel, kuid nagu eelnevast teame, siis statistiliselt on kombineeritud mudel iga vaadatud meetriku korral parem.

Joonis 5 kinnitab kvantitatiivselt varasemates lõikudes tehtud järeldusi: kombineeritud mudel mitte ainult ei paranda üldist täpsust, vaid vähendab ka erindite osakaalu, mis on töö praktilise rakendatavuse seisukohalt kriitilise tähtsusega. TOPz mudelil on sagedamini väga suurte absoluutvigade esinemist. Vastupidiselt sellele on kombineeritud mudeli puhul suurte vigade sagedus oluliselt väiksem — seda näitab histogrammi sabade madalam tihedus väljaspool määratud IQR piire. Seega võimaldab mudelite lineaarne kombineerimine saavutada mitte ainult väiksemat keskmist viga, vaid ka märkimisväärselt usaldusväärsemat üldist ennustusjaotust.



Joonis 6. TOPz ja XGBoost1000 ning kombineeritud mudeli vigade võrdlus.

Joonisel 6 on kujutatud hajuvusdiagramm, mis võrdleb iga andmepunkti puhul TOPz mudeli ja kombineeritud mudeli ζ ennustusvigu. Nagu eelmiselgi joonisel ei ole siin tegu absoluutväärtuste ega keskmiste veamõõdikutega (nagu MAE või MSE), vaid vaadeldakse toorvigu, s.t. erinevusi ennustatud ja tegeliku väärtuse vahel koos märgiga. Vead

võivad seega olla nii positiivsed kui ka negatiivsed, viidates kas üle- või alahindamisele. Punktide tihedus ruumilises jaotuses on esitatud kuusnurkses (hexbin) vormingus, kus värv tähistab andmepunktide arvu igas piirkonnas logaritmilisel skaalal. Tulemuste paremaks tõlgendamiseks on lisatud diagonaaljoon (katkendlik joon), mis tähistab olukorda, kus mõlema mudeli viga on täpselt võrdne.

Diagrammi tõlgendamisel tuleb olla tähelepanelik – kuna vead on esitatud koos märgiga, ei tähenda diagonaalist allapoole jäämine alati paremat tulemust. Täpsemalt, ainult parempoolses ehk positiivsete vigade piirkonnas (kus mõlemad mudelid on vastavat väärtust üle hinnanud) on diagonaalist allpool asuvad punktid näitajaks sellele, et kombineeritud mudel tegi väiksema vea kui TOPz mudel. Vastupidiselt, vasakpoolses ehk negatiivsete vigade piirkonnas (kus ennustus on olnud alahinnatud), tähendavad diagonaalist üleval pool paiknevad punktid, et kombineeritud mudel oli täpsem ehk selle negatiivne viga oli absoluutväärtuses väiksem kui TOPz mudelil.

Graafiku keskosas, kus vead on väiksemad, on näha tihe kontsentratsioon punkte diagonaali läheduses, mis viitab sellele, et enamikus tavapärastest olukordadest käituvad mõlemad mudelid üsna sarnaselt. Siiski on silmatorkav, et paljud punktid asuvad kas positiivsete või negatiivsete väärtuste puhul nendes piirkondades, mis viitavad kombineeritud mudeli paremale täpsusele. Lisaks näeme, et väga suurte vigade puhul, eriti positiivsetel väärtustel, jäävad paljud punktid alla diagonaali, kinnitades, et kombineeritud mudel suudab TOPz mudeli vigu oluliselt leevendada just suurte ennustusvigade korral. Kokkuvõttes toetab see joonis varasemaid järeldusi: mudelite lineaarne kombineerimine vähendab süsteemseid vigu ja aitab saavutada üldiselt väiksemate äärmuste ja väiksema varieeruvusega tulemusi.

5.2 Tõenäosusjaotusel treeniv mudel

Selles peatükis kirjeldan TOPz tõenäosusjaotusel treenivat mudelit. Kuna selles osas on vaja ennustada 500-kohalist vektorit, on mudeli treenimisaeg ligikaudu 50 kuni 500 korda pikem (XGBoost suudab mingil määral arvutusi paralleliseerida) ja seega võtab arvutuste tegemine oluliselt rohkem aega. Seetõttu vaatan selles osas vaid XGBoost mudeleid. Muid parameetreid peale n_enumerators ma ei muuda ja sellel vaatan 4 erinevat väärtust 50, 100, 200 ja 1000. Lisaks proovin ka XGBoost mudelit parameetritega n_enumerators = 100 ja multi_strategy = 'multi_output_tree' ehk mudelit, mis lubab ühel puul kajastada korraga mitut väljundi väärtust. Seda teen ainult ühe mudeli XGBoost100 korral, et seda saaks otseselt võrrelda samaväärse iga vektori väärtust sõltumatult ennustava mudeliga.

Tabel 3. Erinevate masinõppe mudelite võrdlus TOPz mudeliga spektroskoopiliste andmetega võrdlusel

Mudel	**R^2 skoor**	**MAE**	**MSE**	**Erindite osakaal**	**Treeningaeg (s)**
TOPz	0,887	0,0274	0,00209	0,0215	-
XGBoost50	0,877	0,0302	0,00221	0,0250	227
XGBoost100	0,883	0,0290	0,00209	0,0204	610
XGBoost200	0,771	0,0281	0,00208	0,0197	953
XGBoost1000	0,893	0,0271	0,00193	0,0199	3886
XGBoost5000	0,894	0,0271	0,00194	0,0200	11094
XGBoost100 multitree	0,868	0,0313	0,00223	0,0194	821
XGBoost1000 kombineeritud	0,893	0,0271	0,00193	0,0203	3886

Tabelis 3 on esitatud erinevate XGBoost mudelivariantide ning võrdlusmudeli TOPz statistilised jõudlusnäitajad: determinatsioonikordaja (R^2), keskmine absoluutviga (MAE), keskmine ruutviga (MSE), erindite osakaal ning mudelite treenimiseks kulunud aeg sekundites. Antud näitajate alusel on võimalik hinnata iga mudeli ennustustäpsust ning praktilist kasutuskõlblikkust, võttes arvesse nii täpsust kui ka arvutuslikku efektiivsust.

XGBoost100 ja XGBoost100 multitree mudelite võrdlus võimaldab mõista, kuidas mõjutas mudeli struktuurne muutus – nn multitree strateegia – mudeli jõudlust. Mõlemad mudelid on treenitud sama arvu iteratsioonidega (100), kuid nende arhitektuuriline erinevus seisneb selles, et standardne XGBoost kasvatab igas iteratsioonis ühe puu, samas kui multitree strateegia kasvatab igas iteratsioonis mitu puud, mis vastutavad korraga mitme sihtmuutuja (või klassi) ennustamise eest või mille eesmärk on ühes iteratsioonis paremini haarata andmete mitmemõõtmelisi mustreid.

Esmalt uurime, kas parameeter multi output tree teeb mudelit paremaks. Tabelist 3 selgub, et standardne XGBoost100 saavutab paremad tulemused kui XGBoost100 multitree mudel. XGBoost100 puhul on determinatsioonikordaja $R^2 = 0,883$, keskmine absoluutviga MAE = 0,0290 ning keskmine ruutviga MSE = 0,00209. Multitree mudeli vastavad näitajad on halvemad: $R^2 = 0,868$, MAE = 0,0313 ja MSE = 0,00223. Samuti on multitree mudeli treeningaeg (821 s) pikem kui standardmudelil (610 s), mis viitab suuremale arvutuslikule koormusele, kuigi tulemuslikkus on madalam.

Multitree strateegia tööpõhimõtteks on kasvatada ühes boosting-iteratsioonis korraga mitu puud, mitte ainult üks. Klassifitseerimisülesannetes kasutatakse seda lähenemist tihti mitmeklassiliste probleemide puhul, kus iga puu prognoosib eri klassi. Regressiooni puhul võib see strateegia olla kasulik siis, kui mudel püüab samaaegselt haarata erinevaid sihtmuutujate korrelatsioone või sisemisi struktuurseid omadusi. Käesoleval juhul, kus eesmärk on ühe pideva sihtmuutuja täpne ennustamine, ei toonud multitree lähenemine kaasa soovitud täpsuse paranemist, vaid hoopis mõningase halvenemise.

Seega võib järeldada, et antud regressiooniülesande puhul ei ole multitree strateegia sobiv – selle lisakomplekssus ei paku täpsuse paranemist, kuid kasvatab arvutuslikku koormust. Standardne XGBoost100 osutus efektiivsemaks, pakkudes paremat tasakaalu täpsuse ja ressursikasutuse vahel, seega teiste iteratsioonide arvude juures ma enam ei kasuta võrdluseks seda parameetrit.

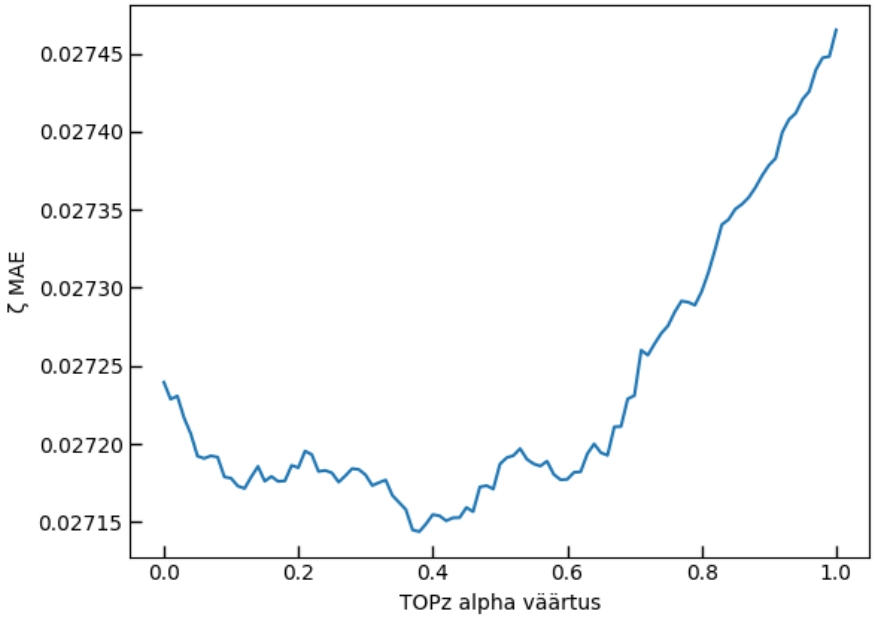
Tabelist 3 ilmneb, et XGBoost1000 mudel saavutab parima üldise tulemuse. Selle determinatsioonikordaja $R^2 = 0,893$ on kõrgeim kõigist testitud mudelitest, viidates väga heale seosele ennustatud ja tegelike väärtuste vahel. Samuti on sellel mudelil üks madalamaid MAE väärtusi (0,0271), mis tähendab, et keskmine ennustusviga on väga väike. Lisaks paistab XGBoost1000 silma madala MSE-ga (0,00193), mis viitab sellele, et ka suuremad ennustusvead esinevad harva ning mudel on stabiilne ja usaldusväärne kogu andmestikus. Väärib märkimist, et madal MSE väärtus viitab paremale võimekusele vältida suuri kõrvalekaldeid, mis on kriitilise tähtsusega näiteks äärmusväärtustele tundlikes rakendustes.

Tabelist 3 ilmneb ka, et erindite osakaalu poolest jääb XGBoost1000 samuti heale tasemele (0,0199), olles märkimisväärselt väiksem võrreldes TOPz mudeliga (0,0215). Kuigi mudel XGBoost200 näitab väiksemat erindite osakaalu (0,0197), ei saa seda pidada paremaks valikuks, kuna selle R^2 väärtus (0,771) on oluliselt madalam ja viitab halvale üldisele sobivusele. Antud juhul madal erindite osakaal ei kompenseeri üldise täpsuse puudujääki. Sama kehtib ka XGBoost100 multitree mudeli kohta.

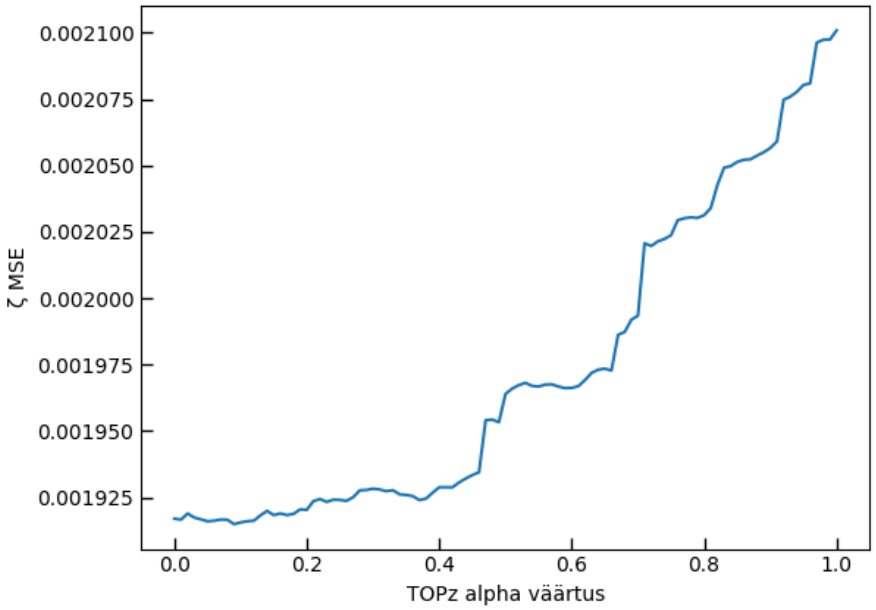
Arvesse tuleb võtta ka treeninguaega. XGBoost1000 mudeli treenimine võtab 3886 sekundit, mis on märgatav ajakulu võrreldes näiteks XGBoost100 (610 s) või XGBoost50 (227 s) mudelitega. Siiski tuleb rõhutada, et lõplik mudel treenitakse tavaliselt vaid üks kord, samas kui selle ennustusi kasutatakse potentsiaalselt väga palju kordi. Seetõttu võib suurema treeningukulu aktsepteerimine olla õigustatud, kui see tagab oluliselt parema täpsuse ja töökindluse. Samas on mudeli XGBoost1000 treeningaeg küllalt lühike, et ka 100 korda suurema andmestiku korral on võimalik mudelit kasutada. XGBoost5000 on küll sarnase täpsusega, aga oluliselt aeglasem kui XGBoost1000.

Kokkuvõttes võib tabeli 3 põhjal järeldada, et XGBoost1000 on kõige tasakaalustatum ja täpsem mudel antud andmekogumiga töötamisel. See tagab väga hea üldistamisvõime, madalad ennustusvead ning väikese hulga erindeid. Kuigi mudeli treening nõuab rohkem

aega, kaalub saavutatud täpsus ja usaldusväärsus selle puuduse üles, mistõttu on see valitud parimaks mudeliks edasistes analüüsides ja rakendustes.



Joonis 7. Kombineeritud mudeli MAE TOPz ja XGBoost logaritmilise kombinatsiooni TOPz kaalust sõltuvana tõenäosusjaotuste mudelil.



Joonis 8. Kombineeritud mudeli MSE TOPz ja XGBoost logaritmilise kombinatsiooni TOPz kaalust sõltuvana tõenäosusjaotuste mudelil.

Järgnevas analüüsis on kasutatud kahe mudeli – XGBoost1000 ja TOPz – prognoositavate väärtuste kombineerimiseks kaalutud geomeetrilist keskmist, eesmärgiga saavutada veelgi täpsem lõppennustus. Geomeetriline keskmine sobib hästi olukordadesse, kus kombineeritakse prognoositavaid tõenäosusjaotusi või mitte-negatiivseid väärtusi, võimaldades mitme allika usaldusväärset ja proportsionaalset sulandamist. Kirjeldan siin geomeetrilise keskmise abil kombineerimist, kuna see andis parema MAE ja MSE väärtused kui lineaarne kombineerimine. Lineaarse kombineerimise parimatel kaaludel MAE ja MSE olid vastavalt 0,0270 ja 0,00197; võrdluseks parimatel kaaludel geomeetrilisel kombineerimisel olid MAE ja MSE vastavalt 0,0271 ja 0,00199. Edasi kirjeldan vaid geomeetrilist kombineerimist. Kombineeritud ennustuse üldkuju on järgmine:

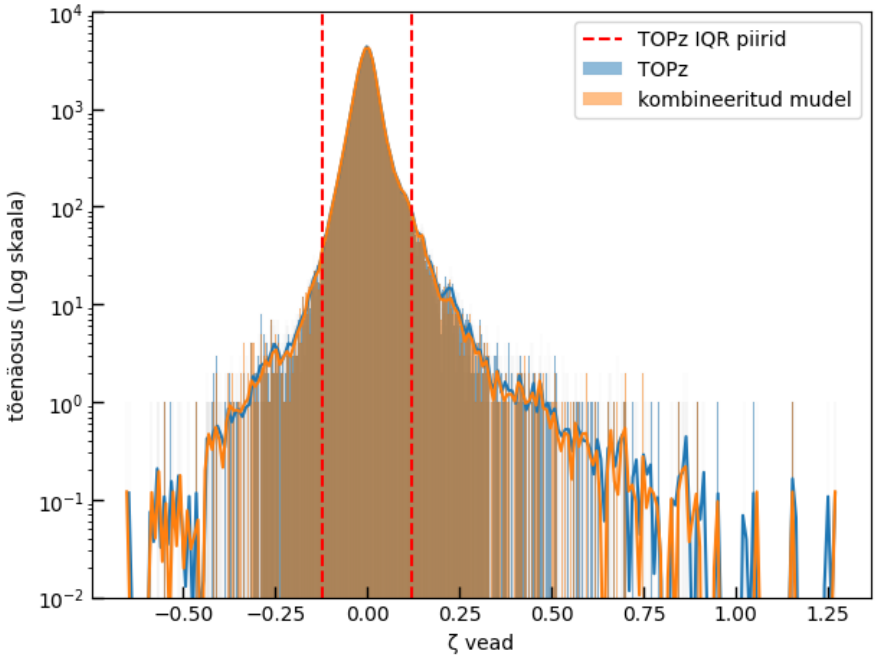
$$\hat{y}_{\text{kombi}} = y_{\text{TOPz}}^{\alpha} \cdot y_{\text{XGB}}^{1-\alpha} \tag{4}$$

kus y_{TOPz} tähistab TOPz mudeli poolt prognoositud vektorit ja y_{XGB} tähistab XGBoost1000 mudeli prognoositud vektorit, millest kõrgeima väärtuse asukoha järgi leian ennustatava ζ, ning $\alpha \in [0, 1]$ on kaalutegur, mis määrab TOPz osakaalu lõpptulemuses. Tegelikus arvutusprotsessis on kasutatud NumPy funktsioone, kus kombineerimine toimub järgmiselt:

$$\text{np.pow}(Y_{\text{TOPz}}, \alpha) \cdot \text{np.pow}(Y_{\text{XGB}}, 1 - \alpha) \tag{5}$$

Et leida optimaalne kaalutegur α, viin läbi süstemaatilise analüüsi erinevate α väärtuste lõikes, mõõtes igal sammul saadud kombineeritud mudeli keskmist absoluutviga ja keskmist ruutviga. Ülaltoodud joonistel on kujutatud MAE ja MSE muutumine vastavalt α väärtusele vahemikus 0-st 1-ni. Jooniselt 7 tuvastati MAE põhjal, et minimaalne viga saavutatakse väärtusel $\alpha = 0,38$, mis tähendab, et optimaalses kombinatsioonis on TOPz aste 0,38 ning XGBoost1000 aste 0,62 lõppennustuse tõenäosusjaotuse leidmisel.

MSE graafik 8 kinnitab samuti, et väiksemad α väärtused annavad üldiselt paremaid tulemusi, viidates sellele, et XGBoost1000 mudeli panus aitab efektiivselt vähendada suurte vigade esinemist. Seetõttu võib järeldada, et kaalutud geomeetriline keskmine ei ole mitte ainult matemaatiliselt sobiv lähenemine kahe mudeli sulandamiseks, vaid annab ka reaalselt mõõdetava täpsusparanduse, kui kasutada õigesti optimeeritud kaale.



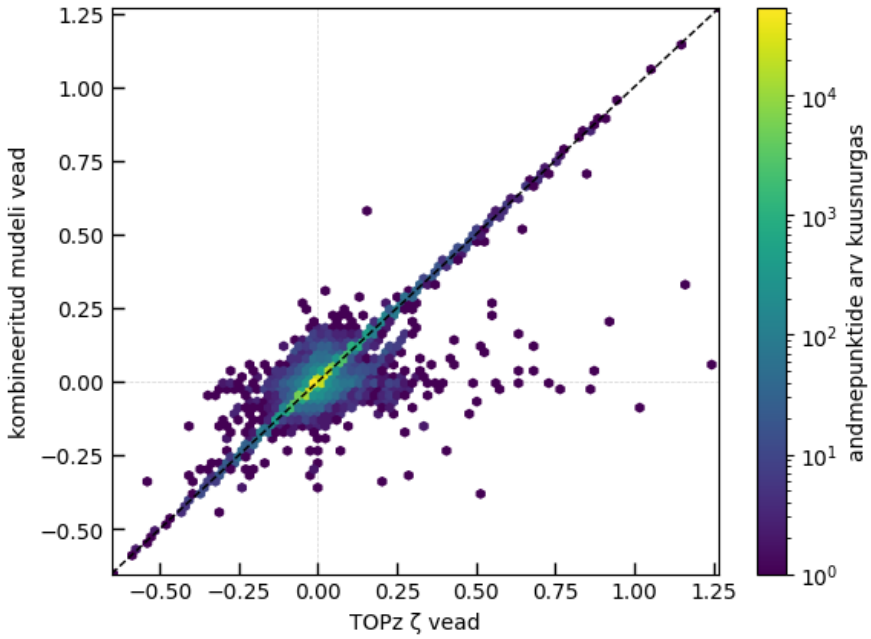
Joonis 9. TOPz ja XGBoost vigade jaotus tõenäosusjaotuste mudelil.

Joonisel 9 on kujutatud TOPz ja kombineeritud mudeli ζ ennustusvigade jaotused logaritmilisel skaalal, kus sinine joon esindab TOPz mudelit ning oranž joon vastab XGBoost1000 ja TOPz mudelite vahel geomeetrilise keskmise põhjal kombineeritud mudelile. Diagrammil on samuti näidatud punase katkestatud joonega TOPz mudeli $3 \cdot$ IQR-põhised piirid, mida kasutatakse varasemalt suurte vigade ehk erindite tuvastamiseks.

Vigade jaotust võrreldes on selgelt näha, et kombineeritud mudeli puhul on jaotus koondunum nulli ümber, mis tähendab, et enamik vigu jääb väiksemaks võrreldes TOPz mudeliga. See ilmneb eriti keskosas, kus kombineeritud mudeli tihedus saavutab suurema tipu, samas kui TOPz mudeli kurv on laiem. Oluline erinevus ilmneb ka jaotuse sabades: kombineeritud mudeli histogramm langeb kiiremini kui TOPz-l, viidates väiksemale suurte vigade ehk erindite sagedusele. Võrreldes joonist 9 joonisega 5, on näha, et

tõenäosusjaotusel treeniv mudel on efektiivsem suuremate vigade parandamisel, aga vigade puhul, mis jäävad $3 \cdot \mathrm{IQR}$ piiridesse, on pigem tõenäoliseimal ζ väärtusel treeniv mudel efektiivsem.

Graafik kinnitab visuaalselt, et kaalutud geomeetriline keskmine kahe mudeli vahel aitab saavutada väiksemate vigadega tulemusi. TOPz mudelil esineb suuremaid kõrvalekaldeid, mis on erindite tuvastamise seisukohalt oluline puudus. Vastupidiselt pakub kombineeritud mudel kitsamat, paremini koondunud jaotust, mis näitab usaldusväärsemat ja stabiilsemat ennustusvõimet. Seega toetab joonis varasemaid kvantitatiivseid tulemusi ja rõhutab kombineeritud lähenemise eelist täpsemate ja järjepidevamate prognooside saavutamisel.



Joonis 10. TOPz ja kombineeritud mudeli ζ ennustamise vigade jaotus võrreldes spektroskoopilise ζ_{spec} väärtusega tõenäosusjaotuste mudelitel.

Joonisel 10 on kujutatud TOPz ja kombineeritud mudeli ζ-vigade omavaheline võrd-

lus hajuvusdiagrammi kujul. Iga punkt graafikul vastab ühe andmepunkti vea väärtusele TOPz mudeli (horisontaaltelg) ja kombineeritud mudeli (vertikaaltelg) korral. Punktide kattumise korral on tihedus esitatud kuusnurkses struktuuris ning värvikaart näitab andmepunktide arvu logaritmilises skaalas. Diagonaaljoon tähistab olukorda, kus mõlema mudeli viga on võrdne – punktid sellel joonel viitavad olukorrale, kus kumbki mudel ei olnud teisega võrreldes parem.

Oluline on tähele panna, et sarnaselt joonisele 6 ei ole antud joonisel kujutatud vigade absoluutväärtusi, vaid toorvead ehk need säilitavad oma märgi. Seetõttu tuleb graafikut tõlgendada järgmiselt: ainult parempoolses ehk positiivsete vigade piirkonnas diagonaaliga samas veerandis (kus mõlemad mudelid on üle hinnanud) tähendavad diagonaalist allpool asuvad punktid, et kombineeritud mudel andis täpsema ennustuse. Vasakpoolses (negatiivsete vigade) piirkonnas diagonaaliga samas veerandis kehtib vastupidine – täpsemad ennustused paiknevad diagonaalist ülalpool. Muudes veerandites on üks mudel ülehinnanud ja teina alahinnanud; nendel juhtudel tuleb hinnata, kumb mudel erineb vähem nullist.

Joonisel 10 on märgata märkimisväärset hulka punkte, mis paiknevad diagonaalist selgelt eemal. See viitab sellele, et kombineeritud mudel ja TOPz annavad mõnel juhul väga erinevaid tulemusi. Seda saab selgitada kombineerimisstrateegiaga: käesolevas lähenemises ei liideta lõplikke ζ väärtusi ega prognoosita tõenäoliseimat asukohta otse, vaid kombineeritakse kahe mudeli tõenäosusjaotused, mille põhjal määratakse tulemuseks saadud jaotuse maksimum. Kui algsed jaotused sisaldavad mitut maksimumi, siis võib kombineerimise tulemusena jääda domineerima TOPz jaotuses olnud madalam maksimum või kujuneda uus hübriidne maksimum. Sellised jaotused on väga levinud, kuna TOPz töötab üksikuid näidiseid sobitades ja võib juhtuda, et mitu füüsikalist näidist on sarnase tõenäosusega. See seletab, miks osa kombineeritud mudeli vigu erineb oluliselt TOPz omadest ja miks mõned punktid paiknevad kaugel diagonaalist — tegemist on

juhtudega, kus jaotuste kombineerimine põhjustas globaalse maksimumi muutumise ning seega suurema vea. Sellised juhtumid on olulised, kuna nad viitavad kombineerimisprotsessi tundlikkusele ning võimaldavad edasisteks täiustusteks hinnata, millal ja miks kombineerimine tulemusi halvendab või parandab. Samuti aitab see tuvastada, millised objektid on võimalik masinõppe järgi paremini ennustada.

Kuidas sellised suured muutused võimalikud on, ilmestab hästi joonis 2, kust on näha, et TOPz ennustab selle objekti jaoks 2 tõenäoliseimat ζ väärtust. Reaalsele spektroskoopilisele ζ väärtusele lähemal Olev maksimum on aga teisest pisut madalam. Samas on XGBoosti järgi õigem maksimum oluliselt kõrge sellest, mis TOPz puhul oli kõrgeim. Seega kombineeritud mudeliga saame maksimumi asukoha, mis on spektroskoopilisele oluliselt lähemal.

Võrreldes joonist 10 joonisega 6 on näha, et tõenäosusjaotuseid kombineerides ei muutu tulemus ühtlaselt nii palju paremaks võrreldes otse ζ ennustamisega. Seda näitab, et joonisel 10 ei teki nii selget diagonaalilähedast suure sagedusega punktide väiksema tõusuga hulka, mis on kaldus kombineeritud mudeli väiksema vea poole, ehk tõenäosusjaotuste mudel ei paranda nii palju keskmiseid tulemusi. Seda näitab ka tabelite 2 ja 3 võrdlus, mis näitab, et otse ζ ennustamine annab väiksema MAE ja MSE 0,0265 ja 0,00166 võrreldes tõenäosusjaotuse mudeli MAE ja MSEga 0,0271 ja 0,00193. Seega annab otse ζ ennustamine täpsemaid tulemusi, aga kui on vaja täiendavalt uurida, kust on vead sisse tulnud, siis tõenäosusjaotust ennustav mudel on paremini interpreteeritav ja annab võimaluse saada põhjalikum ülevaade.

5.3 Edasised sammud

Kuna mudel suutis anda paremaid tulemusi ainult TOPz mudeli väljundist, siis annan oma mudeli koodid Tartu ülikooli TOPz arendava osakonna kätte, et seda saaks tulevikus lisada

TOPz mudeli töövoogu. Seda plaanitakse kasutada juba uute kataloogide ζ väärtuste ennustuste TOPz mudeliga. Täpsemalt on plaan kasutada antud mudelit J-PAS ja WAVES uutel kataloogidel, mis tulevad uute vaatluste tulemusena.

6 Kokkuvõte

Käesolev magistritöö käsitleb masinõppe algoritmide rakendamist fotomeetriliste punanihete (ζ) hindamise täpsustamiseks, keskendudes füüsikalise modelleerimise alusel töötava TOPz algoritmi täiustamisele. Töös kasutati WAVES andmestikku, mis sisaldab nii fotomeetrilisi kui spektroskoopilisi mõõtmisi ning millel põhinevad senised fotomeetrilised punanihete hinnangud on saadud TOPz algoritmiga. TOPz on Tartu Observatooriumis välja töötatud tööriist, mis põhineb spektraalmallide sobitamisel ja Bayesi statistikal ning ei vaja ulatuslikke treeningandmeid. Sellele vaatamata esineb TOPz ennustustes mõõtmismürast või mallide sobimatusest tingitud kõrvalekaldeid ehk erindeid, mida see töö püüdis masinõppe abil parandada.

Töö peamiseks esimeseks eesmärgiks oli leida masinõppemudel, mis suudaks võimalikult täpselt reprodutseerida TOPz algoritmi väljundit, kasutades sisenditena samu fotomeetrilisi andmeid. Selleks hinnati mitmete erinevate masinõppe algoritmide – sealhulgas Lineaarse regressiooni, Random Foresti, Gradient Boostingu ja XGBoosti – võimekust modelleerida TOPz tulemusi. Erinevaid mudeleid võrreldi mitme mõõdiku alusel, sealhulgas determinatsioonikordaja (R^2), keskmine absoluutviga (MAE), keskmine ruutviga (MSE) ning erindite osakaal. Analüüsi tulemusena osutus parimaks XGBoost200 konfiguratsioon, mis andis küll samaväärse tulemuse Random Foresti algoritmiga, aga lisaks pakkus XGBoost märkimisväärselt paremat arvutuslikku efektiivsust, mis muutis selle sobivaks ka suuremate andmestikega töötamiseks.

Edasistes etappides keskenduti TOPz väljundi täiendamisele kahel erineval viisil: esiteks ennustades otseselt ζ väärtust, teiseks modelleerides kogu ζ tõenäosusjaotust. Mõlema lähenemise puhul kombineeriti masinõppemudeli ja TOPz tulemusi, et saavutada täpsem lõppennustus. Kombineerimiseks kasutati lineaarset ja kaalutud geomeetrilist keskmist; viimane osutus tõenäosusjaotuste puhul sobivamaks. Optimaalne kaalutegur

geomeetrilises kombinatsioonis oli $\alpha = 0,38$, mis andis kombineeritud mudelile MAE = 0,0271 ja MSE = 0,00193, selgelt paremad tulemused kui ainult TOPz puhul. Otse ζ väärtuse ennustamine osutus siiski täpsemaks, andes madalama MAE ja MSE (vastavalt 0,0265 ja 0,00166), kuid jaotusepõhine mudel võimaldas paremat visuaalset ja statistilist tõlgendust ning oli väärtuslik keerulisemate juhtumite, näiteks mitme lokaalse maksimumiga jaotuste analüüsis. Seega täiendavad need kaks lähenemist teineteist nii täpsuse kui ka interpreteeritavuse mõttes.

Töös analüüsiti ka kombineerimise mõju vigade jaotusele. Selgus, et kombineeritud mudel andis mitte ainult väiksema keskmise vea, vaid vähendas oluliselt ka suurte vigade ehk erindite osakaalu. Eraldi käsitleti ka juhtumeid, kus kombineeritud mudeli ja TOPz tulemused oluliselt erinevad — näiteks mitme tipu korral jaotuses, kus kombineerimine võib kallutada tulemuse TOPz madalama, kuid füüsikaliselt paremini põhjendatud tõenäosusjaotuse maksimumi suunas. Sellised juhtumid toovad välja, et kombineeritud lähenemine ei anna mitte ainult kvantitatiivselt paremaid tulemusi, vaid aitab ka kvalitatiivselt vältida mõningaid TOPz tüüpilisi vigu.

Kuigi tõenäosusjaotusel põhinev mudel ei ületanud kõikides mõõdikutest tõenäoliseima ζ ennustuse mudelit (nt MAE ja MSE olid veidi kõrgemad), pakkus see suuremat paindlikkust ja interpretatsiooni võimalusi, võimaldades paremini hinnata, kuidas mudel oma otsuseni jõudis. See muudab selle väärtuslikuks tööriistaks juhtumite analüüsil, kus soovitakse mõista, miks üht või teist väärtust eelistati.

Töö praktiliseks tulemuseks on valmis masinõppemudel, mis on sobiv lisand TOPz töövoogu. Mudel võimaldab olemasolevaid hinnanguid täpsustada ilma vajaduseta uute spektraalandmete kogumiseks ning seeläbi vähendab kaudselt vajadust mahukate ja kulukate vaatluste järele. Valminud mudelit on kavas kasutada Tartu Ülikooli teadusgrupi poolt edaspidi TOPz süsteemi osana, sh planeeritud J-PAS ja WAVES kataloogide töötlemisel. Seeläbi annab töö olulise panuse fotomeetrilise punanihke hindamise valdkonda

ning loob võimalused täpsemaks ja usaldusväärsemaks kosmoloogiliseks analüüsiks ka suurandmestike kontekstis.

Viidatud kirjandus

[Baldry, 2018] Baldry, I. K. (2018). Reinventing the slide rule for redshifts: the case for logarithmic wavelength shift.

[Boquien, M. et al., 2019] Boquien, M., Burgarella, D., Roehlly, Y., Buat, V., Ciesla, L., Corre, D., Inoue, A. K., ja Salas, H. (2019). Cigale: a python code investigating galaxy emission. *AA*, 622:A103.

[Brescia et al., 2021] Brescia, M., Cavuoti, S., Razim, O., Amaro, V., Riccio, G., ja Longo, G. (2021). Photometric redshifts with machine learning, lights and shadows on a complex data science use case. *Frontiers in Astronomy and Space Sciences*, Volume 8 - 2021.

[Brescia, M. et al., 2014] Brescia, M., Cavuoti, S., Longo, G., ja De Stefano, V. (2014). A catalogue of photometric redshifts for the sdss-dr9 galaxies. *AA*, 568:A126.

[Calzetti, 2014] Calzetti, D. (2014). The scaling of star formation: from molecular clouds to galaxies. *Proceedings of the International Astronomical Union*, 10(S309):121–128.

[Cavuoti, S. et al., 2012] Cavuoti, S., Brescia, M., Longo, G., ja Mercurio, A. (2012). Photometric redshifts with the quasi newton algorithm (mlpqna) results in the phat1 contest. *AA*, 546:A13.

[Driver et al., 2016] Driver, S. P., Davies, L. J., Meyer, M., Power, C., Robotham, A. S. G., Baldry, I. K., Liske, J., ja Norberg, P. (2016). *The Wide Area VISTA Extra-Galactic Survey (WAVES)*, page 205–214. Springer International Publishing.

[Fotopoulou, S. ja Paltani, S., 2018] Fotopoulou, S. ja Paltani, S. (2018). Cpz: Classification-aided photometric-redshift estimation. *AA*, 619:A14.

[Hatfield et al., 2022] Hatfield, P. W., Jarvis, M. J., Adams, N., Bowler, R. A. A., Häußler, B., ja Duncan, K. J. (2022). Hybrid photometric redshifts for sources in the cosmos and xmm-lss fields. *Monthly Notices of the Royal Astronomical Society*, 513(3):3719–3733.

[Hildebrandt, H. et al., 2010] Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., Assef, R. J., Banerji, M., Benítez, N., Brammer, G. B., Budavári, T., Carliles, S., Coe, D., Dahlen, T., Feldmann, R., Gerdes, D., Gillis, B., Ilbert, O., Kotulla, R., Lahav, O., Li, I. H., Miralles, J.-M., Purger, N., Schmidt, S., ja Singal, J. (2010). Phat: Photo-z accuracy testing. *AA*, 523:A31.

[Hubble, 1929] Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173.

[Karttunen et al., 2007] Karttunen, H., Kröger, P., Oja, H., Poutanen, M., ja Donner, K. (2007). *Fundamental astronomy*. Springer, United States, 5th ed edition.

[Landolt, 2007] Landolt, A. U. (2007). Standardization in the Classical UBVRI Photometric System. In Sterken, C., editor, *The Future of Photometric, Spectrophotometric and Polarimetric Standardization*, volume 364 of *Astronomical Society of the Pacific Conference Series*, page 27.

[Laur, J. et al., 2022] Laur, J., Tempel, E., Tamm, A., Kipper, R., Liivamägi, L. J., Hernán-Caballero, A., Muru, M. M., Chaves-Montero, J., Díaz-García, L. A., Turner, S., Tuvikene, T., Queiroz, C., Bom, C. R., Fernández-Ontiveros, J. A., González Delgado, R. M., Civera, T., Abramo, R., Alcaniz, J., Benítez, N., Bonoli, S., Carneiro, S., Cenarro, J., Cristóbal-Hornillos, D., Dupke, R., Ederoclite, A., López-Sanjuan, C., Marín-Franch, A., de Oliveira, C. M., Moles, M., Sodré, L., Taylor, K., Varela, J., ja Ramió, H. V. (2022). Topz: Photometric redshifts for j-pas. *AA*, 668:A8.

[Masters et al., 2015] Masters, D., Capak, P., Stern, D., Ilbert, O., Salvato, M., Schmidt, S., Longo, G., Rhodes, J., Paltani, S., Mobasher, B., Hoekstra, H., Hildebrandt, H., Coupon, J., Steinhardt, C., Speagle, J., Faisst, A., Kalinich, A., Brodwin, M., Brescia, M., ja Cavuoti, S. (2015). Mapping the galaxy color–redshift relation: Optimal photometric redshift calibration strategies for cosmology surveys. *The Astrophysical Journal*, 813(1):53.

[Merz et al., 2025] Merz, G., Liu, X., Schmidt, S., Malz, A. I., Zhang, T., Branton, D., Burke, C. J., Delucchi, M., Ejjagiri, Y.Š., Kubica, J., Liu, Y., Lynn, O., Oldag, D., ja Collaboration, T. L. D. E.Š. (2025). Deepdisc-photoz: Deep learning-based photometric redshift estimation for rubin lsst. *The Open Journal of Astrophysics*, 8.

[Momtaz et al., 2022] Momtaz, A., Salimi, M. H., ja Shakeri, S. (2022). Estimating the photometric redshifts of galaxies and qsos using regression techniques in machine learning.

[Pathi et al., 2024] Pathi, I. M., Soo, J. Y. H., Wee, M. J., Zakaria, S. N., Ismail, N. A., Baugh, C. M., Manzoni, G., Gaztanaga, E., Castander, F. J., Eriksen, M., Carretero, J., Fernandez, E., Garcia-Bellido, J., Miquel, R., Padilla, C., Renard, P., Sanchez, E., Sevilla-Noarbe, I., ja Tallada-Crespí, P. (2024). Annz+: an enhanced photometric redshift estimation algorithm with applications on the pau survey.

[Sadeh et al., 2016] Sadeh, I., Abdalla, F. B., ja Lahav, O. (2016). Annz2: Photometric redshift and probability distribution function estimation using machine learning. *Publications of the Astronomical Society of the Pacific*, 128(968):104502.

[Savitzky ja Golay, 1964] Savitzky, A. ja Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

[Tempel et al., 2025] Tempel, E., Laur, J., Jones, Z. R., Kipper, R., Liivamägi, L. J., Pandey, D., Sakteos, G., Tamm, A., Triantafyllaki, A. N., ja Tuvikene, T. (2025). Topz: photometric redshifts using template fitting applied to gama survey.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Hannes Kuslap**, 10mmautori nimi

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Masinõppe algoritmide rakendamine füüsikast lähtuvate fotomeetriliste punanihete täpsustamisel,

   mille juhendaja(d) on Elmo Tempel ja Taavi Tuvikene,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3,0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

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

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

Hannes Kuslap

15.05.2025