Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Kristina Muhu

Log-suhte teisendustel põhinevad ennustusmudelid soolevähi diagnoosimiseks mikrobioomi andmetelt

Matemaatika ja statistika õppekava

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendaja: Oliver Aasmets, MSc

Tartu 2021

Log-suhte teisendustel põhinevad ennustusmudelid soolevähi diagnoosimiseks mikrobioomi andmetelt

Magistritöö

Kristina Muhu

Lühikokkuvõte

Magistritöö eesmärk on uurida kompositsionaalsete andmete teooriast tuttavate log-suhte transformatsioonide ja masinõppemeetodite mõju soolevähi ennustusmudelite üldistusvõimele. Täpsemalt uuritakse aditiivse, paariviisilise ning tsentreeritud log-suhte teisenduste ning elastse võrgu ja juhumetsa meetodite kombinatsioonidel põhinevate soolevähki prognoosivate mudelite ennustus- ja üldistusvõimet viie populatsiooni soolestiku mikrobioomi andmeid kasutades. Lisaeesmärk on teada saada parimaid tulemusi andev metoodika, mida saaks võtta aluseks Eesti soolevähi sõeluuringu programmi edendamiseks.

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

Märksõnad: mikrobioom, kompositsioon, log-suhte teisendus, regulariseeritud logistiline regressioon, juhumets

Prediction models based on log-ratio transformations for diagnosing colorectal cancer using microbiome data

Master's thesis

Kristina Muhu

Abstract

The aim of the master's thesis is to study the effect of log-ratio transformations on the accuracy and generalizability of the machine learning models built for colorectal cancer detection. In particular, the predictability and generalizability of colorectal cancer prediction models based on combinations of additive, pairwise and centered log-ratio transformations and machine learning methods such as elastic net and random forest are investigated using fecal microbiome profiles from five different populations. An additional

goal is to find out the methodology that gives the best results, which could be used as a basis for improving the Estonian colorectal cancer screening program.

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

Key Words: microbiome, composition, log-ratio transformation, regularized logistic regression, random forest

Sisukord

Sissejuhatus

Inimesega seotud mikroorganismide ehk mikrobioomi seosed eri haigustega on populaarne uurimisvaldkond nii Eestis kui mujal maailmas. Üks sellistest haigustest, mille puhul on nähtud seosed kõige tugevamad ning mille levimus on kasvamas, on soolevähk. Eestis (täpsemalt Geenivaramus) on soov välja arendada efektiivsed mudelid, millega mikrobioomi ning muude riskitegurite abil diagnoosida soolevähki ilma invasiivsete meditsiiniliste sekkumisteta. Käesolev magistritöö aitab mõista, millised metoodilised lähenemised on selle eesmärgi saavutamiseks just mikrobioomi andmete kasutamisel efektiivsed.

Magistritöö peamine eesmärk on uurida log-suhte transformatsioonide mõju soolevähi tuvastamiseks kasutatavate masinõppemudelite ennustus- ja üldistusvõimele. Töö esimeses osas antakse lühiülevaade inimese mikrobioomist ning selle seosest soolevähiga, sealjuures kirjeldatakse seni soolevähi diagnoosimiseks kasutusel olevaid protseduure. Teises osas kirjeldatakse mikrobioomi andmete olemust ning nende spetsiifilisi omadusi, näiteks kompositionaalsust, mitmemõõtmelisust ja hõredust, ning analüüsimisega kaasnevaid probleeme.

Kolmandas osas antakse ülevaade mikrobioomi koosluse määramise spetsifiikast tulenevast andmete kompositsionaalsusest, sealjuures tuuakse illustreerivaid näiteid kompositsionaalsuse eiramisest tulenevatest probleemidest. Seejärel kirjeldatakse John Aitchisoni välja pakutud põhiprintsiipe kompositsionaalsete andmete analüüsiks. Edasi tutvustatakse erinevaid log-suhte transformatsioone: aditiivne, paariviisiline ja tsentreeritud log-suhte teisendused, mille rakendamine mikrobioomi andmetele lahendab enamik kompositsionaalsusest tekkivaid probleeme.

Lisaks kirjeldatakse töö kolmandas osas kahte mikrobioomi analüüsimisel kasutatavat populaarset masinõppemeetodit: regulariseeritud logistilisel regressioonil põhinevat elastse võrgu (*Elastic Net*) meetodit ning otsustuspuudel põhinevat juhumetsa (*Random Forest*) algoritmi.

Töö neljandas osas antakse ülevaade analüüsimisel kasutatavatest andmetest ning ennustusmudelite loomise tehnilisest poolest. Viiendas osas võrreldakse erinevate teisenduste-algoritmide kombinatsioonide tulemusi soolevähi ennustamisel, kirjeldatakse mudelite üldistatavust andmestikes, kus osasid liike ei ole võimalik tuvastada ning mudelite omadusi. Tulemusi kirjeldavale osale järgneb sisuline arutelu.

Käesolev magistritöö on kirjutatud ja tabelid koostatud kasutades tekstitöötlusprogrammi LaTeX. Töös kasutatud andmete eeltöötlus, kirjeldav analüüs, mudelite treenimine ning illustreerivad joonised on tehtud kasutades rakendustarkvara *R*.

1. Inimese mikrobioom ja seosed soolevähiga

Inimese mikrobioom ehk mikroorganismide genoomide kogum sisaldab üle saja korra rohkem geene ning on seetõttu geneetiliselt oluliselt mitmekesisem ja paindlikum kui inimese genoom. Tänaseks on teada, et mikrobioom suudab toota mitmeid eluks vajalikke molekule, näiteks vitamiine ja lühikese ahelaga aminohappeid, veel enam on mikrobioomil oluline roll inimese immuunsüsteemi arengus. Seejuures on igal inimesel võrdlemisi unikaalne mikrobioom. Kui kahe inimese genoomides on üle 99% sarnasust, siis mikrobioomi puhul on see hinnanguliselt alla 5%. Lisaks eristatakse eri piirkondade mikrobioome (naha, suu, söögitoru, seedetrakti jne), mida mõjutavad mitmesugused tegurid, alustades ravimitest ning lõpetades eluviisiga [1].

Magistritöös keskendutakse soolestiku mikrobioomile, mille seoseid tervise, toitumise ning muude faktoritega on enim uuritud. On teada, et soolestiku mikrobioomi kooslust mõjutab (pikaaegne muutumatu) toidumenüü ning piisavalt äärmuslik lühiajaline muutus toitumises võib põhjustada eri inimeste mikrobioomide üksteisele sarnanemise. Samuti on näidatud, et antibiootikumite varajases eas tarvitamine mõjutab soolestiku mikrobioomi, mis omakorda võib põhjustada rasvumist, astmat, erinevaid soolehaigusi ja muid tervisehäireid. Soolestiku mikrobioomi võib muutuda ka vähese une tõttu. Kuid mikrobioom võib ka ise mõjutada mitmeid füsioloogilisi protsesse. Näiteks on teada, et mikrobioom võib mõjutada leptiini kontsentratsiooni ning selle kaudu ka söögiisusid [1].

Inimese mikrobioomil on oluline osa inimese immuunsüsteemi, neuroloogiliste haiguste ja ka vähi arengus. Üks selliseid haiguseid on soolevähk [2]. Jämesoolevähk (ühisnimetus käärsoole- ja pärasoolevähile), mis saab enamasti alguse (healoomulisest) limaskesta kasvajast, on üks levinumaid surmaga lõppevatest vähitüüpidest maailmas. Eestis haigestub jämesoolevähki ligikaudu 1000 inimest aastas, sealjuures peaaegu võrdselt mehi ja naisi. Kuna jämesoolevähi tekkimise risk suureneb vanuse kasvades ning haiguse varajases staadiumis avastamise korral on seda võimalik ravida, on oluline vähi algusjärgus avastamine ning ravi õigeaegne alustamine [3].

Nii Eestis (alates 2016. aastast) kui ka mujal riikides viiakse jämesoolevähi haigestumise ennetamiseks ja vähendamiseks läbi sõeluuringuprogramme. Sõeluuringu peamise (ja vähi avastamise proaktiivse) meetodina kasutatakse enamasti peitveretesti, mis aitab tuvastada väljaheites leiduvat verd. Positiivse testi saanutel viiakse järgmise sammuna läbi koloskoopia (ehk jämesoole uurimine torukujulise videokaameraga varustatud instrumendiga). Et tegu on äärmiselt invasiivse protseduuriga, loobub märgatav hulk inimesi positiivse peitveretesti järel koloskoopias osalemast. Seetõttu jääb endiselt palju vähijuhte tuvastamata [3].

Uuringud on näidanud olulist seost soolestiku mikrobioomi ja jämesoolevähi vahel. Soolevähi diagnoosiga inimeste ning tervete inimeste väljaheite proove võrreldes on leitud nii soolevähiga seotud baktereid kui ka erinevatele vähistaadiumitele spetsiifilisi mikroorganisme. Seetõttu on välja pakutud, et peitvere (täpsemalt peitvere immuunkeemilise - FIT) testi jaoks võetud väljaheite proove võiks tulevikus kasutada mikrobioomi määramiseks. See annaks võimaluse kombineerida peitveretesti ja soolestiku mikrobioomi soolevähi avastamisel eesmärgiga testi sensitiivsust suurendada ning seeläbi invasiivse koloskoopia protseduuride arvu vähendada [2].

Käesolevas magistritöös pakub huvi soolevähi ennustamine soolestiku mikrobioomi kasutades.

2. Mikrobioomi analüüsimine

Mikrobioom on keerukas ja dünaamiline süsteem, kus paljud mikroorganismid täidavad inimorganismile vajalikke ülesandeid samal ajal üksteisega ressursside ning elukoha eest pidevalt konkureerides. Seda dünaamikat juhivad suurel määral inimeste toitumise- ning elustiiliharjumused, mille tõttu on iga inimese mikrobioom väga personaalne. Nimetatud dünaamika ja isikutevaheline varieeruvus koos andmete kogumisest tulenevate eripäradega muudab mikrobioomi andmetega töötamise väljakutseterohkeks [4].

Selleks, et mikrobioomi edukalt analüüsida, on vaja aru saada andmete tekkemehhanimist, omadustest ning erinevate analüüsimeetoditega kaasnevatest probleemidest. Üks paljulubavamaid kasutusvaldkondi mikrobioomi kaasamisel personaalmeditsiini on mikrobioomi põhjal haiguste diagnoosimine ning prognoosimine. Sarnaselt teiste keerukate bioloogiliste andmete põhjal ennustusmudelite loomisega kasutatakse viimastel aastatel ka mikrobioomi puhul üha enam masinõppel põhinevaid lähenemisi [5].

Näiteks on proovitud mikrobioomi põhjal diagnoosida maksatsirroosi, soolevähki, põletikulisi soolehaigusi, ülekaalulisust ja teist tüüpi diabeeti selliste masinõppemeetoditega nagu juhumets (*random forest*), tugivektorklassifitseerimine (*support vector machine)*, lassoregressioon (LASSO) ja elastne võrk (*elastic net*) [6]. Samuti on populaarne meetod sügavõpe (sügavad närvivõrgud, *deep learning*), mille abil on andmetelt prognoositud näiteks antibiootikumide kasutamisel tekkivaid kõrvalmõjusid [5].

Kuigi masinõppemeetodite kasutamisega on saadud väga hästi haigust ennustavaid mudeleid, on probleemiks loodud mudelite üldistusvõime uutele andmetele. Sageli ühes populatsioonis välja töötatud ennustusmudel ei suuda teises populatsioonis haigeid tervetest eristada või on mudeli prognoosivõime selgelt vähenenud. Üheks võimalikuks põhjuseks on asjaolu, et masinõppemeetodeid mikrobioomi andmetele rakendades ei arvestata mikrobioomi andmete eripäraga [5].

2.1. Ülevaade mikrobioomi andmete tekkest

Mikrobioomi koosluse määramiseks on erinevaid võimalusi. Eesmärk on uurida, millised mikroobid (eelkõige bakterid) uuritavas proovis esinevad ning milline on nende arvukus. Selleks kogutakse keskkonnast (näiteks süljest, väljaheidetest, nahalt) proov, millest eraldatakse kõik seal leiduv DNA. Seejärel DNA sekveneeritakse ehk määratakse DNA-järjestus, millele järgneb bioinformaatiline

analüüs liigikoosluse määramiseks, kus saadud DNA-lugemite andmebaasidega võrdlemise abil saadakse teada vastav liik. Kõige sagedasemad on *shotgun* ja 16S rRNA geeni sekveneerimisel põhinevad metoodikad [7].

Nagu nimigi ütleb, otsitakse 16S rRNA meetodi puhul proovist ainult ühte markerit, täpsemalt 16S rRNA geeni, mis on kõigil bakteritel olemas. Erinevaid liike (perekondi) on võimalik määrata 16S rRNA geeni varieeruvate piirkondade järgi [8, lk 6]. *Shotgun* meetodi puhul ei otsita ainult üht geeni, vaid loetakse kogu geneetiline materjal ning püütakse selle põhjal liike määrata. Viimase meetodiga saadakse enamasti usaldusväärsemad hinnangud bakteriliikide arvukustele [9, lk 33-35].

Mõlema meetodi puhul saadakse tulemuseks tabel, kus on ridades indiviidid, veergudes tuvastatud bakteriliigid ning väärtusteks täisarvud, mis kirjeldavad, mitu korda bakterit proovis detekteeriti. Matemaatiliselt asuvad sellised andmed maatriksis $\mathbf{X}$, millel on m (bakteriliikide arv) veergu ning n (indiviidide arv) rida:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix},$$

kus elemendid $x_{ij} \in \mathbb{N} \cup \{0\}$, $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$ esindavad bakterite määratud arvukusi j-nda bakteriliigi ja i-nda indiviidi kohta [9, lk 35-44].

Maatriksi $\mathbf{X}$ read summeeruvad iga indiviidi korral arvuks, mis sõltub sekveneerimise sügavusest (*sampling depth* või *sequencing depth*) ehk loetud DNA järjestuste arvust inimese kohta. Andmete analüüsimise seisukohalt on oluline, et sekveneerimise sügavus sõltub otseselt sekvenaatorist ehk masinast, millega DNA-d loetakse [9, lk 44]. Seega saadud andmed ei kujuta endast bakterite tegelikke absoluutarve organismis. Piiratud sekveneerimise sügavuse tõttu on bioinformaatilise analüüsi tulemusena saadud bakterite arvud üksteisest sõltuvad ning nad kannavad endas ainult relatiivset informatsiooni, mis annab aluse andmete teisendamiseks suhtelisteks sagedusteks (TSS transformatsioon, *total sum scaling*) [10].

Lisaks mikrobioomi liigilisele kooslusele ollakse sageli huvitatud ka bioloogilistest funktsioonidest, mida mikrobioom on võimeline läbi viima. Sellisel juhul uuritakse liikide määramise asemel mikrobioomile omaseid geene, mille kohta on nende funktsioon teada. Kuid ka nii saadakse tulemuseks andmetabel, kus ridades on indiviidid ja veergudes kas geenid või bioloogilised funktsioonid ning saadud andmed kannavad ainult suhtelist informatsiooni [8, lk 30].

2.2. Mikrobioomi andmete omadused

Mikrobioomi andmetel on mitmed spetsiifilised omadused, mis põhjustavad väljakutseid andmete analüüsimisel ning prognoosimudelite loomisel.

1. Mikrobioomi andmed on **kompositsionaalsed** (kompositsionaalsetest andmetest täpsemalt peatükis 3.1.). Ühe indiviidi (proovi) loenduste koguarv sõltub sekvenaatori võimekusest, mille tõttu tekivad sõltuvused bakteriliikide vaadeldud arvukuste vahel [9, lk 56]. Et sekveneerimissügavus ei kanna analüüsimisel olulist infot, saab mikrobioomi analüüsimiseks kasutada ainult bakterite suhtelisi arvukusi [8, lk 34].

2. Mikrobioomi andmetele on iseloomulik **mitmemõõtmelisus**. Olenevalt huvi pakkuvast infost sisaldavad mikrobioomi andmed sadu erinevaid bakteriliike, tuhandeid bioloogilisi funktsioone või miljoneid geene, kuid kordades vähem vaatluseid. Näiteks selles töös kasutatavas mikrobioomi andmestikus on 849 bakteriliiki ning 575 vaatlust. Sellise kõrge dimensionaalsuse tõttu tekib nn $n < p$ (kus p on vaadeldud tunnuste arv) probleem, mis raskendab andmete analüüsimist ning prognoosimudelite loomist [8, lk 34].

3. Mikrobioomi andmetele on omane **hõredus** (*sparse data*) ehk suure hulga nullide esinemine. Kui maatriksi $\mathbf{X}$ element $x_{ij} = 0$, siis sellisel juhul ei ole j-ndat bakteriliiki i-ndas proovis tuvastatud. Nulle eristatakse tekkemehhanismide põhjal järgmiselt [9, lk 73-74]:

   (a) loendus nullid (*sampling/count zeros*) tekivad valimi suuruse või kasutatud tehnikate tõttu;

   (b) ümardatud nullid (*rounded zeros*) tekivad, nagu nimigi ütleb, väärtuste ümardamisest: kui bakteriliikide arvukused andmestikus jäävad alla mingi ümardamisvea või tuvastamispiiri, siis ümardatakse väärtused nulliks;

   (c) sisulised nullid (*essential/structural zeros*) tähistavad tegelikke nulle ehk juhte, kus bakterit proovis ei leidu.

   Mikrobioomi andmetes esinevad nullid ning nendega korrektne tegelemine andmeid analüüsides on oluline uurimisvaldkond, millele on alles hiljuti hakatud tähelepanu pöörama [11].

4. Mikrobioomi andmetel on **hierarhiline struktuur**. Bakteriliigid on evolutsiooniliselt seotud ning neid seoseid liikide vahel saab esitada fülogeneesipuuna. Bakteriliigid, mis on puus

üksteisele lähemal, on teineteisele bioloogiliselt sarnasemad, reageerivad keskkonnateguritele üldiselt samalaadselt ning täidavad sarnaseid funktsioone [8, lk 30].

5. Mikrobioomi andmed on keeruka **kovariatsiooni ja korrelatsiooni struktuuridega**. Bakteriliigid interakteeruvad inimorganismis mitmeti. Enamasti on erinevad liigid pidevas konkurentsis toitainete ja eluruumi pärast, kuid esineb ka liikidevahelisi kommensialistlikke seoseid, kus näiteks üks liik toitub teise liigi jääkainetest teda kahjustamata. Et bakterid on omavahel süsteemselt seotud, ei saa vaadelda baktereid sõltumatute organismidena [12].

Osad klassikalised masinõppe algoritmid võimaldavad loetletud omadustest toime tulla vaid mõnega, lassoregressioon näiteks $n < p$ probleemiga, kuid mitte kõigiga. Seetõttu on tähelepanu hakatud pöörama mikrobioomi andmete omadusi arvesse võtvate meetodite arendamisele [13]. Samas on näidatud, et klassikaliste algoritmide ennustustäpsust saab parandada, kui võtta arvesse tulemusi kompositsionaalsete andmete teooriast [14].

3. Kasutatud metoodika

Järgnevalt antakse ülevaade kompositsionaalsete andmete olemusest ja omadustest ning kaasnevatest probleemidest. Seejärel kirjeldatakse kompositsionaalsete andmete analüüsimiseks kasutatavaid log-suhte teisendusi. Viimaks antakse ülevaade kahest mikrobioomi andmetel enimkasutatavast masinõppemeetodist: elastne võrk ning juhumets.

3.1. Kompositsionaalsed andmed

On selge, et mikrobioomi andmete puhul saab vaadelda ainult liikide suhtelisi arvukusi, st mikrobioomi andmed on kompositsionaalsed.

Definitsioon 1. *Vektor $\vec{x} = (x_1, x_2, \ldots, x_D)$ ($D \in \mathbb{N}$) defineerib D-osalise kompositsiooni, kui kõik komponendid $x_1, \ldots, x_D$ on rangelt positiivsed reaalarvud ning kannavad ainult relatiivset informatsiooni.*

Kompositionaalsed vektorid paiknevad hüpertasandil, täpsemalt simpleksil, mis on defineeritud järgmiselt:

$$S^D = \left\{ \vec{x} = (x_1, x_2, \ldots, x_D) \mid x_i > 0,\ i = 1, 2, \ldots, D;\ \sum_{i=1}^{D} x_i = \kappa \right\}, \tag{1}$$

kus κ on suvaline positiivne konstant. Tavapäraselt $\kappa = 1$ või $\kappa = 100$, mis tähendab, et elemendid $x_1, \ldots, x_D$ tähistavad osakaale või protsente [15, lk 5].

Kompositsionaalsete andmete peamine probleem seisneb selles, et andmepunktid ei asetse eukleidilises ruumis, vaid hoopis Aitchisoni simpleksil S^D. Eirates andmete kompositsionaalsust, võivad analüüsi käigus saadud tulemused olla väärad. Näiteks võivad kompositsionaalsust eirates tekkida mittetõesed (*spurious*) korrelatsioonid – vaatamata bakteritevahelisele tegelikule seosele tekivad summakitsenduse tõttu kompositsiooni komponentide vahele negatiivsed korrelatsioonid.

Olgu näiteks $\vec{X} = (X_1, \ldots, X_D)$, $D \in \mathbb{N}$, selline juhuslik vektor, mille korral $\sum_{i=1}^{D} X_i = 1$ (*). Siis kovariatsiooni omadusi kasutades saab kirjutada:

$$\mathrm{Cov}(X_1, \sum_{i=1}^{D} X_i) = \mathrm{Cov}(X_1, X_1) + \cdots + \mathrm{Cov}(X_1, X_D).$$

Samas (*) tõttu kehtib ka

$$\mathrm{Cov}(X_1, \sum_{i=1}^{D} X_i) = \mathrm{Cov}(X_1, 1) = 0,$$

ning seega

$$\mathrm{Cov}(X_1, X_1) + \cdots + \mathrm{Cov}(X_1, X_D) = 0,$$

millest

$$-D(X_1) = \mathrm{Cov}(X_1, X_2) + \cdots + \mathrm{Cov}(X_1, X_D). \tag{2}$$

Selgub, et välja arvatud juhul, kui X_1 on konstantne, on võrduse 2 vasak pool negatiivne, mistõttu vähemalt üks kovariatsioonidest peab olema negatiivne. Seega esineb andmestikus negatiivseid korrelatsioone olenemata sellest, kas bakterite arvukused on ka tegelikult omavahel korreleeritud (*negative correlation bias*) [9, lk 60-61].

Veel enam, alamkompositsiooni (vaadeldakse $\vec{X}$ alamhulka, nt $\vec{X}' = (X_1, X_2, X_3)$) elementide vahelised Pearsoni korrelatsioonid võivad esialgse kompositsiooni elementide vahelistest korrelatsioonidest erineda nii väärtuse suuruse kui ka märgi poolest [9, lk 61]. Olukorda, kus alamkompositsioonil tehtavad järeldused erinevad tervel kompositsoonil tehtutest, nimetatakse alamkompositionaalseks invariantsuseks (*subcompositional incoherence*). Eelnevalt kirjeldatu tõstatab probleemi, mille tõttu ei saa kompositsionaalsetele andmetele otse rakendada standardseid korrelatsioonil põhinevaid tehnikaid, nt korrelatsioon- või peakomponentide analüüsi.

Lisaks raskendab andmete kompositsionaalsus uuritavates gruppides erinevate arvukustega bakterite tuvastamist. Klassikalised lähenemised nagu t-test võivad kompositsionaalsete andmete puhul suurendada esimest liiki vea tõenäosust ehk tõsta valepositiivsete leidude arvu [9, lk 59-60].

3.1.1. Näide probleemist kompositsionaalsete andmetega

Olgu nelja bakteri B_1, B_2, B_3, B_4 tegelikud arvukused kolme patsiendi mikrobioomi proovis sellised, et bakterit B_4 esineb kõikidel indiviididel sama palju, bakteriliike B_3 ja B_2 on kahel esimesel võrdselt ning B_1 arvukus on sama esimesel ja kolmandal indiviidil (tabelis 1).

Tabel 1. Bakterite tegelikud arvukused (n) ja proportsioonid (p) patsientide mikrobioomi proovides.

Proovid	**Bakteriliigid**							
	B_1		B_2		B_3		B_4	
	n	p (%)	n	p (%)	n	p (%)	n	p (%)
Indiviid 1	40	13	60	20	110	37	90	30
Indiviid 2	80	24	60	18	110	32	90	26
Indiviid 3	40	19	70	33	10	5	90	43

Mikrobioomi andmete puhul saab proovis hinnata aga ainult liikide proportsioone. Vaadeldes nüüd liikide proportsioone, on näha, et näiteks bakteri B_4 protsendid on kõikide indiviidide proovides erinevad, mis võib viia uuritava tunnuse kohta valede järelduste tegemiseni. Näiteks, olgu teada, et indiviidid 1 ja 2 on terved, kuid indiviid 3 on haige. Vaatamata sellele, et tegelikkuses on bakteri B_4 arvukused kõikidel indiviididel samad, on bakteri B_4 proportsioon haige indiviidi korral märgatavalt kõrgem kui teistel patsientidel. Seega võib raviotsuse tegemine proportsioonide põhjal olla vale.

Tabel 2. Bakterite B_1, B_2, B_3 tegelikud arvukused (n) ja proportsioonid (p) patsientide mikrobioomi proovides.

Proovid	**Bakteriliigid**					
	B_1		B_2		B_3	
	n	p (%)	n	p (%)	n	p (%)
Indiviid 1	40	19	60	29	110	52
Indiviid 2	80	32	60	24	110	44
Indiviid 3	40	33	70	58	10	8

Olgu nüüd proovides tuvastatud ainult esimesed kolm bakteriliiki (ehk vaadeldakse alamkompositsiooni) (tabel 2). Lihtsuse mõttes olgu bakterite suhtelised sagedused esialgses kompositsioonis tähistatud (analoogiliselt peatükis 2. esitatule) liigi B_1 korral $\vec{x}_1 = (x_{11}, x_{21}, x_{31})'$ ja B_2 korral $\vec{x}_2 = (x_{12}, x_{22}, x_{32})'$ ning analoogiliselt alamkompositsioonis B_1 korral $\vec{y}_1 = (y_{11}, y_{21}, y_{31})'$ ja B_2 korral $\vec{y}_2 = (y_{12}, y_{22}, y_{32})'$, kus näiteks element x_{31} vastab tabelis 1 bakteri B_1 arvukusele või proportsioonile kolmanda indiviidi korral. Vaadeldes bakteriliikide B_1 ja B_2 **proportsioonide** vahelisi kovariatsioone kahel juhul (arvestades et $N = 3$), kui

1. kõik neli liiki on tuvastatud (tabel 1 põhjal), kus $\bar{\vec{x}}_1 \approx 19$ ja $\bar{\vec{x}}_2 \approx 24$:

$$\begin{aligned}\mathrm{Cov}_{\vec{x}_1,\vec{x}_2} &= \frac{\sum_{i=1}^{n}\left(x_{i1}-\bar{\vec{x}}_1\right)\left(x_{i2}-\bar{\vec{x}}_2\right)}{N-1} = \\ &= \frac{\left(13-\bar{\vec{x}}_1\right)\cdot\left(20-\bar{\vec{x}}_2\right)+\cdots+\left(19-\bar{\vec{x}}_1\right)\cdot\left(33-\bar{\vec{x}}_2\right)}{2} \approx -0.03,\end{aligned}$$

2. ainult kolm liiki on tuvastatud (tabel 2 põhjal), kus $\bar{\vec{y}}_1 \approx 28$ ja $\bar{\vec{y}}_2 \approx 37$:

$$\begin{aligned}\mathrm{Cov}_{\vec{y}_1,\vec{y}_2} &= \frac{\sum_{i=1}^{n}\left(y_{i1}-\bar{\vec{y}}_1\right)\left(y_{i2}-\bar{\vec{y}}_2\right)}{N-1} = \\ &= \frac{\left(19-\bar{\vec{y}}_1\right)\cdot\left(29-\bar{\vec{y}}_2\right)+\cdots+\left(33-\bar{\vec{y}}_1\right)\cdot\left(58-\bar{\vec{y}}_2\right)}{2} \approx 0{,}63,\end{aligned}$$

on näha, et nii kovariatsioonide suurused kui ka märgid erinevad teineteisest ning tabeli 2 (ehk alamkompositsiooni) põhjal tehtavad järeldused erineksid tabeli 1 (tervel kompositsioonil) tehtavatest järeldustest.

3.1.2. Põhiprintsiibid kompositsionaalsete andmete analüüsimisel

Šoti matemaatik/statistik John Aitchison avaldas 1986. aastal senini statistika raudvarasse kuuluva raamatu „The Statistical Analysis of Compositional Data“, milles kirjeldas peamised põhimõtted kompositsionaalsete andmete analüüsiks.

Kolm põhiprintsiipi tuginevad kompositsionaalsete andmete definitsioonil ning on järgmised:

- **skaala invariantsus** – statistiliste järelduste tegemine kompositsionaalsete andmete kohta ei tohi sõltuda nende skaalast (κ valemis 1), st vektorid $x = (2, 3, 6)$, $y = (20, 30, 60)$ ja $z = (400, 600, 1200)$ esindavad kõik sama kompositsiooni ja on ekvivalentsed, sest $10x = y$, $20x = z$, $20y = z$ [15, lk 7-9];

- **permutatsiooni invariantsus** – statistiliste järelduste tegemine kompositsionaalsete andmete kohta ei tohi sõltuda komponentide järjekorrast vektoris, st ei ole oluline, milline komponent valitakse „esimeseks“, milline „teiseks“ jne [15, lk 9-10];

- **alamkompositsiooni invariantsus** – esiteks peavad alamkompositsioonil tehtud järeldused olema kooskõlas tervel kompositsioonil tehtud järeldustega ning teiseks ei tohi analüüsi

tulemused sõltuda nendest elementidest, mis ei ole alamkompositsioonis (ehk ei tohi sõltuda kaasamata komponentidest) [8, lk 335].

Näites 3.1.1. selgus, et suhteliste sageduste puhul ei ole täidetud alamkompositsiooni invariantsuse nõue, sest alamkompositsioonilt arvutatud korrelatsioonid ei ühtinud tervelt kompositsioonilt arvutatutega. Kui kirjeldatud kolm printsiipi oleksid rahuldatud, siis ei tekiks eelnevalt välja toodud probleeme. Kuna mikrobioomi puhul ei saada metodoloogiliste piirangute tõttu kunagi kätte kõiki proovis leiduvaid bakteriliike, siis tegeletakse alati alamkompositsioonidega, mille tõttu on alamkompositsiooni invariantsuse printsiibi täitmine väga oluline [9, lk 61].

3.2. Log-suhte analüüs

Selleks, et kompositsionaalsetele andmetele standardseid statistilisi meetodeid rakendada, pakkus Aitchison lahendusena välja andmete teisendamise reaalarvude ruumi kasutades nn log-suhte transformatsioone. Nende põhimõte seisneb selles, et kompositsiooni kahe või enama elemendi arvukused või funktsioonid jagatakse omavahel ning saadud jagatisest võetakse logaritm. Kompositsiooni elementide (või elementide sobivate funktsioonide) jagamine tagab skaala invariantsuse ning jagatise logaritmiline aitab lahti saada mittenegatiivsuse kitsendusest [8, lk 336].

Aitchison (ja teised) töötasid arvestades kompositsionaalsete andmete eripära välja mitmed erinevate omadustega log-suhte transformatsioonid: aditiivne log-suhte *(additive log-ratio*, ALR*)*, paariviisilise log-suhte (*pair-wise log-ratio*, PWLR*)* ning tsentreeritud log-suhte (*centred log-ratio*, CLR*)* teisendused. Hilisemalt on välja töötatud mitmeid teisi log-suhte transformatsioone, nagu isomeetriline log-suhte (*isometric log-ratio*, ILR*)* transformatsioon [16], summeeritud log-suhte transformatsioonid [17] ning palju teisi. Töös keskendutakse nendest kolmele esimesele transformatsioonile, mida järgnevalt lähemalt tutvustatakse.

3.2.1. Aditiivne log-suhte transformatsioon

Lihtsaim log-suhte transformatsioon ning Aitchisoni esimene väljatöötatud meetod kompositsionaalsete andmete analüüsiks on aditiivne log-suhte transformatsioon (ALR), mis on defineeritud järgmiselt:

$$\mathrm{alr}(\vec{x}) = \left(\ln\left(\frac{x_1}{x_D}\right), \ldots, \ln\left(\frac{x_i}{x_D}\right), \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right), \tag{3}$$

kus $\vec{x} = (x_1, \ldots, x_D)$ on kompositsioon. Selline transformatsioon teisendab D-elemendilise Aitchison'i simpleksi $D - 1$ dimensionaalseks eukleidiliseks vektoriks, kusjuures valemis 3 murru nimetajas olevaks võrdlus-elemendiks võib võtta suvalise komponendi x_D vektorist $\vec{x}$. Saadud andmetele saab edasi rakendada kõiki standardseid statistilisi meetodeid, mis ei põhine kaugusel [8].

Tabel 3. ALR transformatsioon (valem 3) rakendatud bakterite tegelikele arvukustele (n) ja proportsioonidele (p) (1) kasutades referents-elemendiks liiki B_4.

Proovid	**Teisendused**					
	$\ln(B_1/B_4)$		$\ln(B_2/B_4)$		$\ln(B_3/B_4)$	
	n	p (‰)	n	p (‰)	n	p (‰)
Indiviid 1	−0,81	−0,81	−0,41	−0,41	0,20	0,20
Indiviid 2	−0,12	−0,12	−0,41	−0,41	0,20	0,20
Indiviid 3	−0,81	−0,81	−0,25	−0,25	−2,20	−2,20

ALR transformatsiooni tulemused sõltuvad sellest, millist referents-elementi murru nimetajas kasutatakse. Rakendades ALR transformatsiooni (valem 3) alapeatüki 3.1.1. näiteandmetele (tabel 1) ning valides esimesel juhul võrdluselemendiks sellise liigi, mille arvukus on kõikidel indiviididel sama suur (liik B_4), on tabelis 3 näha, et transformeerimise tulemusena saadud tunnuste väärtused on indiviidide lõikes samad arvutatuna nii tegelikel arvukustel kui ka proportsioonidel. Näiteks leitakse tabel 1 põhjal esimese indiviidi bakterite B_1 ja B_4 proportsioonide põhjal ALR-teisendus järgmiselt:

$$\ln \frac{x_{11}}{x_{14}} = \ln \frac{13}{30} \approx -0,81.$$

Kasutades referents-elemendiks aga sellist liiki, mille arvukus erineb indiviididel (nt liik B_1), on selge, et saadavad tulemused (vt tabel 4) erinevad tabelis 3 esitatud tulemustest. Seetõttu ei ole ALR transformatsiooni tulemused üheselt määratud, vaid sõltuvad väga suurel määral sellest, milline võrdlus-element valitakse.

Tabel 4. ALR transformatsioon (valem 3) rakendatud bakterite tegelikele arvukustele (n) ja proportsioonidele (p) (1) kasutades referents-elemendiks liiki B_1.

Proovid	**Teisendused**					
	$\ln(B_2/B_1)$		$\ln(B_3/B_1)$		$\ln(B_4/B_1)$	
	n	p (%)	n	p (%)	n	p (%)
Indiviid 1	0,41	0,41	1,01	1,01	0,81	0,81
Indiviid 2	−0,29	−0,29	0,32	0,32	0,12	0,12
Indiviid 3	0,59	0,59	−1,39	−1,39	0,81	0,81

ALR transformatsioon rahuldab kõiki Aitchisoni põhiprintsiipe (vt peatükk 3.1.2.). On selge, et ühe teisendatud tunnuse, nt $\ln(B_1/B_4)$ väärtus ei sõltu teiste (mis ei ole valitud referent-elemendiks, nt B_3) bakteriliikide olemasolust ja arvukusest, seega alamkompositsiooni invariantsuse printsiip on täidetud.

ALR transformatsioon on asümmeetriline kompositsiooni elementide suhtes, mille tõttu elementide vahelised kaugused eukleidilises ruumis sõltuvad nimetaja valikust. See omakorda tähendab, et ALR transformatsiooni tulemusel saadud andmeid ei saa analüüsida standardsete kaugusmaatriksil põhinevate statistiliste meetoditega [8, lk 338-339]. Küll aga on praktikas ALR transformatsiooni tulemusel saadud andmeid lihtne tõlgendada, seda juhul kui on suudetud valida õige võrdlus-element [18].

Kui on teada, et nimetajas oleva bakteri tegelik arvukus on muutumatu, siis saab muutuste korral log-suhte väärtustes teha järeldusi lugejas oleva bakteri arvukuse muutuse kohta. Kui nimetaja kohta ei ole teada, kas tegelik arvukus muutub või mitte, siis ei saa ka lugeja muutuse kohta midagi väita. Samas ei ole kunagi täpselt teada, millise bakteri tegelik arvukus ei muutu ning seetõttu on ALR transformatsioon leidnud praktikas pigem vähe kasutust [18].

Selleks, et eespool kirjeldatud nimetaja valiku probleemist üle saada, võib vaadelda log-suhte transformatsiooni, kus kasutatakse nimetajas kõiki bakteriliike – selleks on paariviisiline log-suhte transformatsioon.

3.2.2. Paariviisiline log-suhte transformatsioon

Paariviisiline log-suhte transformatsioon (PWLR) on defineeritud läbi kõikvõimalike vektori $\vec{x}$ elementidevaheliste log-suhete:

$$\text{pwlr}(\vec{x}) = \left(\ln x^*_{ij}; i < j = 1, 2, \ldots, D\right), \text{ kus } x^*_{ij} = \frac{x_i}{x_j}. \tag{4}$$

Kuna kehtib $\ln(A/B) = -\ln(B/A)$, siis piisab logaritmiline suhe arvutada vaid juhtudel, kus $i < j,\ i, j = 1, \ldots, D$. Seega saadakse D- elemendilisest kompositsionaalsest vektorist $D \cdot (D-1)/2$-dimensiooniline eukleidilises ruumis asuv vektor [19].

Tabelis 1 toodud näiteandmetele rakendatud PWLR transformatsiooni (valem 4) tulemused on esitatud tabelis 5. Kuna $D = 4$, siis on elementide arv teisendatud vektoris $4 \cdot (4-1)/2 = 6$.

Tabel 5. PWLR transformatsioon (valem 4) rakendatud bakterite tegelikele arvukustele (n) ja proportsioonidele (p) tabeli 1 põhjal.

Proovid	**Teisendused**					
	$\ln(B_1/B_2)$		$\ln(B_1/B_3)$		...	$\ln(B_3/B_4)$	
	n	p (‰)	n	p (‰)	...	n	p (‰)
Indiviid 1	−0,41	−0,41	−1,01	−1,01	...	0,20	0,20
Indiviid 2	0,29	0,29	−0,32	−0,32	...	0,20	0,20
Indiviid 3	−0,56	−0,56	1,39	1,39	...	−2,20	−2,20

Analoogliselt ALR tulemustele on tabelist 5 näha, et teisenduste tulemus ei sõltu sellest, kas kasutatakse arvukusi või proportsioone. Sarnaselt ALR-teisendusega on ka PWLR-teisenduse korral kõik kolm põhiprintsiipi täidetud.

3.2.3. Tsentreeritud log-suhte transformatsioon

Selleks, et vältida konkreetse referent-elemendi valikut (nagu ALR-teisenduse puhul) ning vältida saadava teisenduse elementide dimensiooni kasvu (nagu PWLR-teisenduse puhul), pakkus Aitchison välja CLR-teisenduse.

Tsentreeritud log-suhte (CLR) transformatsioon on defineeritud järgmiselt:

$$\text{clr}(\vec{x}) = \left(\ln\left(\frac{x_1}{g_m(\vec{x})}\right), \ldots, \ln\left(\frac{x_i}{g_m(\vec{x})}\right), \ldots, \ln\left(\frac{x_D}{g_m(\vec{x})}\right)\right), \tag{5}$$

kus $g_m(\vec{x}) = \sqrt[D]{x_1 \cdot x_2 \cdots x_D}$ on vektori $\vec{x}$ elementide geomeetriline keskmine. Vektori $\vec{x}$ elementide läbi jagamine geomeetrilise keskmisega $g_m(\vec{x})$ viib selleni, et $\mathrm{clr}(\vec{x})$ elementide summa on null ning väärtused jäävad nulli ümber [8].

Kasutades taaskord alapeatüki 3.1.1. näiteandmeid, saab arvutada bakteriliikide arvukustele ning proportsioonidele vastavad geomeetrilised keskmised, näiteks esimesel indiviidil (protsentidega)

$$g_{m_1}(\vec{x}) = \sqrt[4]{13 \cdot 20 \cdot 37 \cdot 30} \approx 23{,}18$$

ning seejärel rakendada CLR transformatsiooni (valem 5).

Tabel 6. CLR transformatsioon (valem 5) rakendatud bakterite tegelikele arvukustele (n) ja proportsioonidele (p) tabeli 1 põhjal, kus g_m on geomeetriline keskmine, mis arvutatakse iga indiviidi jaoks eraldi.

Proovid	**Teisendused**							
	$\ln(B_1/g_m)$		$\ln(B_2/g_m)$		$\ln(B_3/g_m)$		$\ln(B_4/g_m)$	
	n	p (‰)	n	p (‰)	n	p (‰)	n	p (‰)
Indiviid 1	−0,56	−0,56	−0,15	−0,15	0,45	0,45	0,25	0,25
Indiviid 2	−0,04	−0,04	−0,32	−0,32	0,28	0,28	0,08	0,08
Indiviid 3	0,004	0,004	0,56	0,56	−1,38	−1,38	0,81	0,81

Tabelist 6 on näha, et CLR-teisenduse tulemused on taaskord samad nii arvukuste kui proportsioonide korral. Lisaks on saadud elementide summad (nii arvukuste kui protsentide korral) kõikide proovide puhul nullid, nt indiviid 1 korral saame arvukusi kokku liites $-0{,}53+(-0{,}15)+0{,}45+0{,}25 \approx 0$.

CLR-teisenduse puhul ei ole täidetud alamkompositsiooni invariantsuse nõue. Näiteks vaadeldes tabelis 2 toodud alamkompositsiooni andmeid ning arvutades nende põhjal uued tulemused (tabelis 7), selgub et kõikide bakteriliikide CLR-väärtused on võrreldes tabel 7 väärtustega muutunud.

Tabel 7. CLR transformatsioon (valem 5) rakendatud bakterite tegelikele arvukustele (n) ja proportsioonidele (p) tabeli 2 põhjal, kus g_m on geomeetriline keskmine, mis arvutatakse iga indiviidi jaoks eraldi.

Proovid	**Teisendused**					
	$\ln(B_1/g_m)$		$\ln(B_2/g_m)$		$\ln(B_3/g_m)$	
	n	p (‰)	n	p (‰)	n	p (‰)
Indiviid 1	−0,47	−0,47	−0,07	−0,07	0,54	0,54
Indiviid 2	−0,01	−0,01	−0,30	−0,30	0,31	0,31
Indiviid 3	0,28	0,28	0,84	0,84	−1,11	−1,11

CLR transformatsiooni eeliseks on see, et referents ehk murru nimetaja on fikseeritud ning CLR transformeeritud andmed sobivad kasutamiseks kaugusel põhinevate meetodite puhul. Lisaks alamkompositsiooni invariantsuse printsiibi mitterahuldamisele on CLR transformatsiooni puuduseks asjaolu, et kovariatsiooni maatriks ei ole pööratav, mille tõttu on mõne meetodi rakendamine teisendatud andmetele raskendatud [8, lk 338-339].

3.2.4. Probleemid nullidega

Alajaotuses 2.2. selgus, et mikrobioomi andmed on lünklikud ehk andmed sisaldavad palju nulle. See on aga kõikide log-suhte transformatsioonide korral suur probleem, sest logaritm nullist ei ole defineeritud. Et andmestikus olevatel nullidel on erinevad tekkemehhanismid, kasutatakse kompositsionaalsetes andmetes esinevate eri tüüpi nullidega arvestamiseks mitmesuguseid lähenemisi [8, lk 339].

Ümardatud nulle käsitletakse mikrobioomi andmetes enamasti MNAR-tüüpi (*Missing Not At Random*) juhtumitena. Sellised nullid imputeeritakse kasutades väikesi mittenullilisi väärtusi. Selleks asendatakse nullid väikeste konstantidega (*pseudocount*) või kasutatakse erinevaid mitteparameetrilisi ja parameetrilisi meetodeid, nt EM-algoritmi [8, lk 339-340].

Loendus nullidega tegelemiseks kasutatakse praktikas samuti nullide konstandiga asendamist, kui mehhanismist tulenevalt on välja pakutud ka mitmeid Bayesi meetodeid. Sellised meetodid säilitavad vektori elementide esialgsed proportsioonid ning summa kitsenduse, kuid moonutavad siiski mõnevõrra elementidevahelisi seoseid. Kuna kompositsionaalse vektori keskmine on võrdne selle geomeetrilise keskmisega, siis kasutatakse nullide asendamiseks ka GBM (*Geometric Bayesian-Multiplicative*) meetodit, mis säilitab esialgsed komponentidevahelised suhted. Olenemata nende

meetodite headest tulemustest (eriti GBM), ei ole Bayesi ega GBM põhised meetodid kooskõlas skaala invariantsuse kriteeriumiga ning on seetõttu praktikas (töötades mikrobioomi andmetega) vähekasutatavad [8, lk 340].

Sisuliste nullide esinemine kompositsionaalsetes andmetes on kõige keerukam probleem. Sellist tüüpi nullidega tegelemiseks ei ole üldist lihtsat meetodit ning nende asendamine väikese mitte-nullilise arvuga (nagu ümardatud ja loendus nullide korral) ei ole sobilik. Kui selgub, et andmetes võivad olla struktuursed nullid, siis rakendatakse andmetele näiteks Poisson-log normaaljaotuse põhiseid mudeleid [8, lk 340-341]. Paraku on sisuliste nullide esinemise tuvastamine praeguste meetoditega peaaegu võimatu.

Mikrobioomi analüüsimisel ei keskenduta tavapäraselt nullide probleemile väga palju. Kuna üldiselt ei ole piisavalt informatsiooni, et olla kindel millist tüüpi nullidega on mikrobioomi andmestikus tegemist, siis käsitletakse andmetes olevaid nulle loendus nullidena ning nullid asendatakse väikeste positiivsete konstantidega [9, lk 74].

3.3. Masinõppemeetodid

Masinõppemeetodeid on tänaseks edukalt kasutatud erinevate bioloogiliste andmete analüüsimisel, näiteks haiguste diagnostikaks ning haiguste tekke prognoosimiseks. Mikrobioomi andmetel on mitmeid sarnaseid omadusi nimetatud bioloogiliste andmestikega, näiteks n<p probleem ning keerukad korrelatsioonistruktuurid andmetes. Paraku on vähe uuritud, kuidas mõjutab mikrobioomi andmete kompositsionaalsus masinõppemeetodite ennustus- ning üldistusvõimet, sest enamasti on sisendina kasutatud bakterite suhtelisi arvukusi. Järgnevalt tutvustatakse kahte mikrobioomi puhul enamkasutatud masinõppe algoritmi, mille ennustusvõimet töö praktilises osas erinevate andmeteisenduste puhul uuritakse [5].

3.3.1. Regressioonil põhinevad meetodid

Eesmärk on koostada mudel diagnoosimaks soolevähki (väärtusega 1, kui inimesel on haigus, ja väärtusega 0, kui haigust ei ole) kasutades selleks mikrobioomi andmeid. Seega on uuritav tunnus $\vec{y} = (y_1, \ldots, y_n)$ binaarne ($y_i = 1$ või $y_i = 0$, $i = 1, \ldots, n$) ning seletavaid tunnuseid $\mathbf{X} = (\vec{x}_1, \ldots, \vec{x}_p)$ on p arv. Sellisel juhul võib uuritava tunnuse y_i modelleerimiseks kasutada logistilist regressiooni:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tag{6}$$

kus $\pi_i = P(y_i = 1|\mathbf{X})$ ning parameetrite $\beta_0, \ldots, \beta_p$ hinnangud leitakse enamasti suurima tõepära või vähimruutude meetodil [20, lk 135].

Kuna π_i saab esitada ka kujul

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}},$$

siis parameetrite $\beta_0, \ldots, \beta_p$ hinnangud saadakse maksimeerides log-tõepärafunktsiooni [21]:

$$l\left(\beta_0, \ldots, \beta_p\right) = \sum_{i=1}^{n} \left(y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) - \log\left(1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}\right)\right). \quad (7)$$

Mikrobioomi andmetele on iseloomulik valimi suuruse kohta rohke seletavate tunnuste arv. Sellisel juhul tekib aga probleem vähimruutude hinnangute leidmisel, sest vähimruutude hinnangud ei ole $n < p$ korral üheselt määratud [20, lk 218]. Sellisele probleemile on lahenduseks välja pakutud regulariseerimine (*shrinkage*), mille korral parameetrite hinnangud (võrreldes vähimruutude hinnangutega) surutakse nulli poole. Sealjuures on eesmärgiks saada hea prognoosimudel, mille jaoks loobutakse parameetrite hinnangute nihketuse nõudest [20, lk 203].

Populaarseimad meetodid on kant- ja lassoregressioon (vastavalt inglise keeles *Ridge* ja LASSO *regression*) ning nende kombinatsioon elastne võrk (*elastic net*). Nende kolme meetodi korral saab mudelisse lisada korraga kõik p seletavat tunnust, mis on mikrobioomi keerukat korrelatsioonistruktuuri arvestades arvestatav eelis logistilise regressiooni ees [22].

Kantregressiooni korral leitakse parameetrite $\beta_0, \ldots, \beta_p$ hinnangud maksimeerides funktsiooni [21]:

$$l_\lambda^R\left(\beta_0, \ldots, \beta_p\right) = l\left(\beta_0, \ldots, \beta_p\right) - \lambda \sum_{j=1}^{p} \beta_j^2, \quad (8)$$

kus $\lambda \geq 0$ on karistusparameeter ning $\lambda \sum_j \beta_j^2$ on karistusliige.

Kantregressiooni eeliseks on arvutusmahukus. Maksimeerides parameetrite $\beta_0, \ldots, \beta_p$ hinnangute leidmiseks funktsiooni (7) tuleb läbi käia 2^p mudelit, samas fikseeritud λ korral on sobivaks kantregressiooni mudeliks täpselt üks mudel. Kantregressiooni korral lähenevad karistusparameetri λ kasvades parameetrite hinnangud nullile, kuid mitte kunagi täpselt nulliks. Seega on kantregressiooni puuduseks asjaolu, et lõplikku mudelisse jäävad kõik seletavad tunnused, mille tõttu on mudeli interpreteerimine raskendatud [21].

Eelnevalt kirjeldatud probleemist saab jagu **lassoregressioon**, mille korral leitakse parameetrite

$\beta_0, \ldots, \beta_p$ hinnangud maksimeerides funktsiooni [21]:

$$l_\lambda^L\left(\beta_0, \ldots, \beta_p\right) = \left(\beta_0, \ldots, \beta_p\right) - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{9}$$

kus karitusliige funktsioonis 8 on asendatud liikmega $\lambda \sum_j |\beta_j|$.

Piisavalt suure karituselemendi λ tõttu võrdsustatakse osa lassoregressiooni parameetrite hinnangutest $\hat{\beta}_1, \ldots, \hat{\beta}_p$ täpselt nulliga. Seega jääb lõplikku mudelisse võrreldes kantregressiooniga vähem liikmeid, mille tõttu on mudeli tõlgendamine lihtsam. Selle omaduse tõttu on lassoregressioon ka üheks enamlevinud algoritmiks mikrobioomi andmete analüüsimisel, sest lisaks uuritava tunnuse modelleerimisele valib mudel automaatselt välja uuritava tunnuse ennustamiseks olulised mikroobid [21]. Lassoregressioon saab paremini hakkama (prognooside täpsuse mõttes) olukorras, kus väheste seletavate tunnuste kordajad on prognoosimiseks olulised. Kantregressioon aga olukorras, kus seletavate tunnuste kordajad on ligikaudu samas suurusjärgus [21].

Nii kantregressioonil kui ka lassoregressioonil on mõlemal puuduseid ja eeliseid, kuid karistusliikmeid kombineerides on võimalik mudeli omadusi parandada. Sellist mudelit nimetatakse **elastseks võrguks** ning selle karistusliige on kujul

$$\lambda P_\alpha(\beta_0, \vec{\beta}) = \lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right), \tag{10}$$

kus $\alpha \in [0, 1]$, $\vec{\beta} = \left(\beta_1, \ldots, \beta_p\right)$ ning $P_\alpha(\beta_0, \vec{\beta})$ on kombinatsioon kant- ja lassoregressiooni karistusliikmetest [22].

On ilmne, et kui $\alpha = 0$, siis saadakse valemis 10 kantregressiooni karistusliige:

$$\lambda \sum_{j=1}^{p} \left(0\beta_j^2 + (1-0)|\beta_j|\right) = \lambda \sum_{j=1}^{p} |\beta_j|,$$

kui aga $\alpha = 1$, siis lassoregressiooni karistusliige:

$$\lambda \sum_{j=1}^{p} \left(1\beta_j^2 + (1-1)|\beta_j|\right) = \lambda \sum_{j=1}^{p} \beta_j^2.$$

Mikrobioomi andmete puhul on oluliseks aspektiks tunnuste (bakteriliikide) omavaheline korreleeritus, mis võib olla tingitud andmete kompositsionaalsusest (täpsemalt alajaotuses 3.1.). Kantregressiooni korral jäävad mudelisse kõik tunnused, seega ka omavahel tugevalt korreleeritud

tunnused. Näiteks m identse/omavahel tugevalt korreleeritud tunnuse korral saavad kõik m tunnust endale võrdse $1/m$-osa ühe tunnusega sobitatud mudeli kordajast. Lassoregressioon valib mudelisse aga ühe omavahel korreleeritud tunnustest ning jätab teised mudelist välja. Elastne võrk saab hästi hakkama ekstreemsete korrelatsioonidega nö keskmistades korreleeritud tunnused ning valides mudelisse just selle [22].

3.3.2. Juhumets

Järgnev alapeatükk põhineb raamatu „An Introduction to Statistical Learning with Applications in R“ lehekülgedel 303 – 321 [20], kui ei ole märgitud teisiti.

Selleks, et kirjeldada juhumetsa algoritmi, tuleb alustada otsustuspuudest, mida saab kasutada nii regressiooni kui ka klassifitseerimise ülesannete korral. Analoogiliselt peatükis 3.3.1. esitatule on eesmärgiks määrata tunnuse $\vec{y}$ väärtus (0 või 1) tunnuste $\mathbf{X}$ põhjal. Otsustuspuu põhiidee on jagada $\mathbf{X}$ väärtused lõikumatuteks alamhulkadeks $A_1, \ldots, A_J$ ning hulka $A_j, j \in 1, \ldots, J$ kuuluva vaatluse väärtuseks ennustatakse alamhulgas A_j mudeli treenimiseks kasutatud andmete põhjal domineeriv uuritava tunnuse klass.

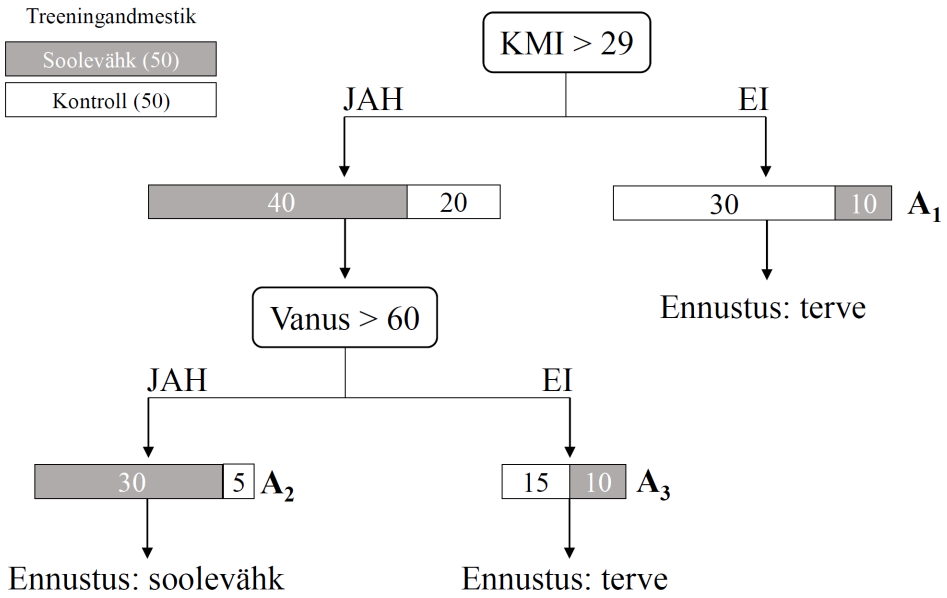
Otsustuspuu meetod lahendab sellise ülesande kahendpuul rekursiivselt vaatluseid konkreetse kriteeriumi järgi kaheks jagades. Täpsemalt tekib tagurpidi puu struktuur: juursõlmes jagatakse vaatlused mingi tunnuse (valitud selline, mis kirjeldab kõige paremini uuritava tunnuse varieerumist) järgi kaheks. Seega esimese sammuna valitakse ennustajaks välja tunnus x_j, mille jaoks leitakse lävend s, mis jagab ennustajate maatriksi $\mathbf{X}$ kaheks regiooniks $\{\mathbf{X}|x_j < s\}$ ning $\{\mathbf{X}|x_j \geq s\}$. Klassifitseerimispuu korral kasutatakse iga tunnuse x_j ja lävendi s hindamiseks Gini indeksit:

$$G = \sum_{k=1}^{K} \hat{p}_{lk}(1 - \hat{p}_{lk}),$$

kus $\hat{p}_{lk}$ tähistab k-ndast klassist pärit treeningvaatluste osakaalu hinnangut l-ndas regioonis, $k \in \{0, 1\}$ ja $l \in \{1, 2, \ldots, J\}$. Gini indeks mõõdab koguvarieeruvust üle K klassi ning väike Gini indeksi väärtus näitab, et leht sisaldab vaatlusi, mille puhul üks uuritava tunnuse klass domineerib.

Sarnaselt valitakse järgmistes tekkinud sõlmedes tunnused, mille järgi gruppides olevad vaatlused kaheks jagada. Niimoodi jätkatakse kuni uuritava tunnuse kirjeldamiseks olulisi tunnuseid rohkem ei leidu või on täidetud algoritmi tööd lõpetav kriteerium, milleks üldiselt on minimaalne vaatluste arv ühes regioonidest $A_j, j \in 1, \ldots, k$. Klassifitseerimispuu tulemuseks saab iga leht (ehk vaatluste grupp) kõige sagedasema uuritava tunnuse y väärtuse. Saadud lehtede põhjal võib prognoosida

tunnuse y klasse ka muude valimite jaoks.



Joonis 1. Näide klassifitseerimispuust, mille abil prognoositakse patsientidel soolevähi olemasolu kasutades kehamassiindeksi (KMI) ja vanuse tunnuseid.

Joonisel 1 on illustreeritud kahe sõlmega klassifitseerimispuud, mis on koostatud soolevähi ennustamise jaoks. Treeningandmestikus on 50 kontrollrühma kuuluvat ning 50 soolevähiga patsienti (kokku 100 patsienti). Otsustuspuu esimesse sõlme on valitud kehamassiindeksi tunnus (KMI). Kriteeriumi {KMI $>$ 29} järgi jagatakse vaatlused kahte gruppi: kui tingimus ei ole täidetud, siis ennustatakse patsient terveks (regioon A_1). Kui tingimus on täidetud, siis tekib uus sõlm, milles on vaatluste kaheks jagamise kriteeriumiks {vanus $>$ 60}. Need patsiendid, kelle vanus on kõrgem kui 60 ennustatakse haigeks (regioon A_2) ning need, kelle vanus on 60 või madalam, saavad prognoosiks terve (regioon A_3). Seega regiooni A_1 kuuluvad need, kelle {KMI $\leq$ 29}, regiooni A_2 kuuluvad need, kelle {KMI $>$ 29} ja {vanus $>$ 60} ning regiooni A_3 need, kelle {KMI $>$ 29} ja {vanus $\leq$ 60}. Vastavalt regioonides kuuluvate treeningvaatluste domineerivale klassile ennustatakse regioonidesse A_1 ja A_3 kuuluvatele vaatlustele, et nad on terved ning regiooni A_2 kuuluvatele, et neil on soolevähk.

Otsustuspuudel on mitmeid häid omadusi kuid ka puuduseid. Tulemusi on lihtne interpreteerida, selgitada ning graafiliselt esitada. Samas ei ole tulemused kuigi täpsed (võrreldes võimalike tulemustega kasutades muid meetodeid, nt lineaarne regressioon vms) ega stabiilsed (lõplikud tulemused on mõjutatud isegi väiksemast muutusest andmestikus) ning puid on kerge ülesobitada.

Kirjeldatud puudusi, eriti tulemuste ennustuse täpsust, on võimalik parandada kasutades juhumetsa (*Random Forest*) algoritmi. Juhumets koosneb sisuliselt juhuslikest otsustuspuudest. Otsustuspuude sõltumatus tekitatakse kahte moodi. Esiteks tekitab juhuslikkuse andmestikust korduvate juhuslike valimite võtmine (*bootstrap aggregation / bagging*).

Olgu $Z_1, \ldots, Z_n$ üksteisest sõltumatud tunnused ning iga tunnuse dispersioon oli σ^2. Siis tunnuste keskmise $\bar{Z}$ dispersioon on σ^2/n. Selleks et vähendada kõrget dispersiooni ning (seeläbi) tõsta täpsust on loomulik võtta populatsioonist B erinevat treeningandmestikku, arvutada nende jaoks prognoosid $\hat{f}^1(x), \ldots, \hat{f}^B(x)$ ning leida nende keskmine

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x).$$

Selline lähenemine ei ole aga praktiline. Seetõttu võetakse B juhuslikku valimit ühest treeningandmestikust ning arvutatakse iga juhusliku valimi jaoks $\hat{f}^{*b}(x),\ b = 1, \ldots, B$. Seejärel leitakse arvutatud prognooside keskmine

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x).$$

Klassifitseerimise ülesande korral toimub enamushääletus, kus ühe vaatluse jaoks leitakse prognoos kõigi B valimite korral ning enim esinenud klass valitaksegi lõpp-prognoosiks.

Teiseks tekitatakse juhumetsa algoritmi puhul otsustuspuude juhuslikkus mudeli ehitamiseks kasutatud seletatavate tunnuste juhusliku valikuga. Kui üks seletavatest tunnustest kirjeldab uuritavat tunnust teistest rohkem, siis enamik genereeritavatest juhuslikest puudest kasutab juursõlmes vaatluste jagamiseks just seda seletavat tunnust. Seetõttu on tekkinud puud üksteisele sarnased ning saadud prognoosid omavahel tugevalt korreleeritud.

Selleks et vältida sarnaste puude tekkimist kaalutakse iga otsustuspuu iga sõlme loomise jaoks ainult t juhuslikku seletavat tunnust kõigi p seletava tunnuse seast ($t < p$). Selliselt võetakse arvesse ka uuritavat tunnust vähem kirjeldavad tunnused.

Seega juhumetsa algoritmi puhul valitakse treeningandmestikust B arv juhuslikke valimeid. Iga valimi pealt ehitatakse otsustuspuu selliselt, et rekursiivselt korratakse järgnevaid samme (1-3) seni, kuni on saavutatud minimaalne soovitud vaatluste arv sõlmes:

1. valitakse juhuslikud t tunnust kõikide seletavate tunnuste seast,
2. t tunnuse seast valitakse parim tunnus, mille järgi vaatlused gruppidesse jagada,

3. valitud tunnuse järgi jagatakse puu alamharudesse.

Iga sellise juhusliku otsustuspuu korral arvutatakse prognoos ning seejärel võetakse üle kõigi otsustuspuude vastu enamusotsus [23, 588].

Võrreldes ühe puu ehitamisega on juhuslike puude tulemused täpsemad, kuid neid on keerukam interpreteerida, näiteks ei saa tulemusi graafiliselt puu kujul esitada ning ei ole lihtne aru saada, millised tunnused on uuritava tunnuse kirjeldamiseks olulisemad.

3.3.3. Mudelite treenimine

Masinõppealgoritmidel on mitmeid parameetreid, mis mõjutavad mudeli lõplikku kuju ning mudelite ennustustäpsust. Selleks, et leida optimaalne mudel, millel oleks hea ennustusvõime ning mis üldistuks hästi ka uutele andmetele, tuleb leida selleks sobivad mudeli parameetrid. Mudelite „sobitamiseks“ ehk parameetrite valikuks kasutatakse enamasti ristvalideerimist, täpsemalt k-kordset ristvalideerimist [20, lk 182-183].

Üldiselt jagatakse mudeli üldistatavuse hindamiseks kasutatav andmestik kaheks: teening- ja testandmestikuks. Kasutades sobivat algoritmi ehitatakse mudel treeningandmetele ehk treenitakse mudel. Seejärel rakendatakse treenitud mudelit testandmetele ning hinnatakse mudeli ennustustäpsust. k-kordse ristvalideerimise puhul jagatakse vaatlused juhuslikult võrdse suurusega k-ks grupiks, mida kasutatakse korda mööda testandmestikuna. Enamasti kasutatakse praktikas $k = 5, 10, n$. Iga k korral sobitatakse mudel ülejäänud treeningandmetel, arvutatakse $\mathrm{AUC}_i,\ i = 1, \ldots, k$, andmestikust eemaldatud (test) vaatluste pealt ning leitakse ristvalideerimiseviga [20, lk 182-183]:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{AUC}_i,$$

kus AUC on ROC-kõvera alune pindala, mis on enimkasutatav mõõdik klassifitseerimismudelite headuse hindamiseks.

Mudeli parameetrite valimiseks hinnatakse mitmeid erinevaid parameetrite kombinatsioone. Iga parameetrite kombinatsiooni jaoks leitakse ristvalideerimisviga $CV_{(k)}$ ning sobivaks parameetri väärtuseks valitakse väikseima ristvalideerimisveaga kombinatsioon. Selliselt on võimalik leida optimaalseid parameetreid iga algoritmi jaoks. Näiteks, kõigi kolme peatükis 3.3.1. kirjeldatud regulariseeritud logistilise regressiooni meetodi korral leitakse sobiv λ või α valemites (8), (9) ja (10) parameetreid sobitades [20, lk 227]. Juhumetsa puhul on sobitamist vajavateks parameetriteks

näiteks igasse sõlme juhuslikult valitud tunnuste arv ning stopp-kriteerim otsustuspuude ehitamiseks. Viimaks hinnatakse mudel kogu treeningandmestikul uuesti kasutades ristvalideerimise teel leitud parameetrite väärtusi [20, lk 321].

3.4. Log-suhte teisenduste mõju mudelite ennustustäpsusele kirjanduses

Siiani ei ole praktikas kuigi palju uuritud log-suhte transformatsioonide mõju mikrobioomi andmeid kasutatavate masinõppemeetodite prognoosivõimele. Järgnevalt kirjeldatakse kirjandusest tulemusi, mis on saadud erinevate masinõppealgoritmide ning log-suhte transformatsioonide kombineerimisel kompositsionaalsetel andmetel.

Tolosana-Delgado jt uurisid isomeetrilise log-suhte transformatsiooni (ILR) ning CLR ja PWLR transformatsioonide mõju juhumetsa ennustusvõimele eesmärgiga prognoosida sügaval maa sisemuses olevate suurte maakooreplokkide omavahelist kokkupuudet Austraalias kasutades maapinna geokeemilisi andmeid. Uuringus näidati, et väga hästi töötab juhumetsa ning PWLR kombinatsioon, sealjuures paranes mudelite ennustusvõime kõigi log-suhte transformatsioonide kasutamisel võrreldes suhtelisi sagedusi kasutatavate mudelitega. Lisaks selgus, et mudeli täpsus tõuseb, kui kaasata kompositsiooni rohkem elemente, sealjuures saadakse paremaid tulemusi kasutades PWLR teisendust [14].

Zhang ja Shi uurisid samuti erinevate log-suhte teisenduste (ALR, ILR, CLR) ja masinõppemeetodite kombinatsioonide prognoosimisvõimet, küll aga ennustades mulla osakeste suurusklasse. Täpsemalt uuriti K-lähima naabri (*K-nearest neighbor*), mitmekihilise närvivõrgu (*multilayer percepron neural network*), juhumetsa, tugivektorklassifitseerimise (*support vector machines*) ja *extreme gradient boosting* meetodeid. Leiti, et üldiselt saadakse võrreldes suhteliste sageduste kasutamisega paremad tulemused kasutades log-suhte transformatsioone. Samuti näidati, et juhumets töötab nii regressiooni- kui klassifitseerimisülesande korral kõikidest vaadeldud meetoditest paremini [24].

Quinn ja Erb näitasid (kasutades inimese eri piirkondade mikrobioomi andmeid ennustamaks erinevaid haiguseid), et ka regulariseeritud logistilise regressiooni (täpsemalt LASSO) puhul töötab mudel paremini kui bakterite suhtelisi sageduste asemel kasutada CLR-transformeeritud andmeid [25].

Eelnevalt välja toodud tulemusi on keeruline laiendada magistrtöös käsitletavale probleemile, mille

eesmärgiks on soolevähi ennustamine mikrobioomi andmetelt. Üldiselt tuleb masinõppemeetodite puhul siiski lähtuda probleemi ning andmete spetsiifikast [7]. Siiski on eelnimetatud tulemused heaks aluseks log-suhte transformatsioonide kasutamise jaoks.

4. Andmete analüüs

Töö praktilises osas kasutatakse mikrobioomi andmestikku, mis on koostatud Wirbeli jt poolt läbi viidud uuringus, mille tulemused on avaldatud artiklis „Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer“. Artiklis kasutati Saksamaal kogutud proovide ning nelja varem publitseeritud uuringu (USA, Austria, Hiina, Prantsusmaa) *shotgun* meetodil põhinevaid väljaheidete mikrobioomi (metagenoomide) andmeid soolevähi ennustusmudelite loomiseks [26].

Kõikide riikide proovid on kogutud enne vähiraviga alustamist. Väljaheite proovid on võetud Prantsusmaa ja Austria korral enne ning Saksamaa, USA ja Jaapani korral peale koloskoopia läbi viimist. Veel tasub märkida, et ka proovide säilitamine on riigiti erinev: näiteks enne DNA eraldamist ja sekveneerimist säilitati USA proove külmutatult vähemalt 25 aastat. Lisaks erineb populatsiooniti ka DNA eraldamise protseduur [26].

4.1. Kirjeldav statistika

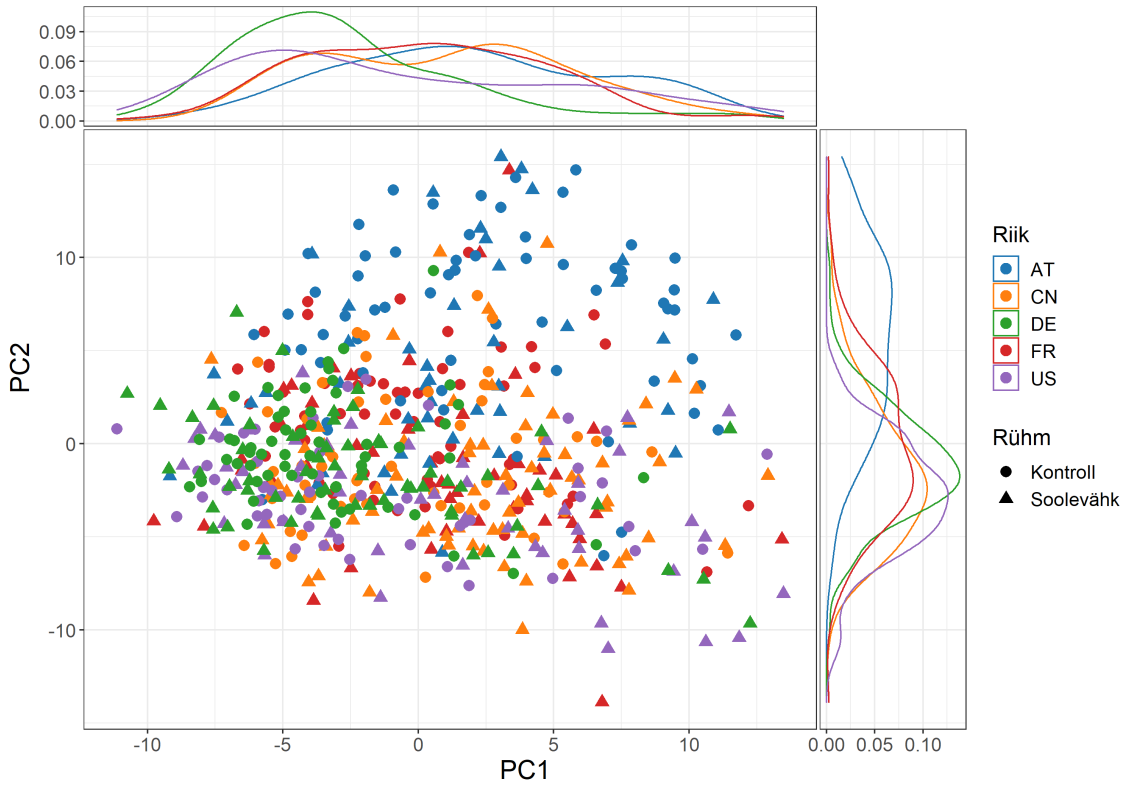
Andmestikus on olulisemad tunnused vanus, sugu, kehamassiindeks (KMI), andmete päritoluriik ning soolevähi diagnoos (soolevähi- või kontrollrühm). Kasutada on 575 indiviidi andmed, täpsemalt 285 soolevähiga vaatlust ning 290 kontrollrühma kuuluvat vaatlust. Soolevähi juhtude ning kontrollide jaotus riikide lõikes näidatud tabelis 8. Kõige rohkem indiviide on Hiina ning kõige vähem USA valimis, vastavalt 128 ja 104. Andmestikus on bakteriliike 849.

Tabel 8. Valimite suurused (n) riikide lõikes soolevähi diagnoosiga ja tervete rühmades.

Riik	**Haiged**	**Terved**	**Kokku**
	n	n	n
Austria	46	63	109
Hiina	74	54	128
Prantsusmaa	53	61	114
Saksamaa	60	60	120
USA	52	52	104
Kokku	**285**	**290**	**575**

Keskmine kehamassiindeks (KMI) ning indiviidide vanus, mis on ühtlasi soolevähi riskifaktoriteks, erinevad mõningal määral riigiti (lisas joonistel 10, 11). Austria andmetes on keskmine KMI nii kontroll kui soolevähi rühmas märgatavalt kõrgem kui teistes riikides. Lisaks on keskmine kehamassiindeks Hiina, Saksamaa ja Prantsusmaa korral soolevähi rühmas kõrgem kui kontrollrühmas. Samuti on keskmine vanus Austria valimis kõrgem kui teistes riikides. Tähele tasub panna ka seda, et soolevähi rühmas on vanus kõikides riikides kõrgem kui kontrollrühmas.

Selleks, et kirjeldada populatsioonidevahelist erinevust mikrobioomi kooslustes, tehakse CLR transformatsiooniga teisendatud mikrobioomi andmestikule peakomponentanaüüsi. CLR-teisenduse rakendamine on kirjeldatud alapeatükis 4.2.1. Andmete transformeerimine CLR-teisenduse abil enne peakomponentanalüüsi aitab toime tulla kompositsionaalsusest tulenevate piirangutega [27]. Joonisel 2 on esitatud proovide jaotumine esimese kahe peakomponendi järgi.



Joonis 2. Mikrobioomi CLR-teisendusega andmete peakomponentanalüüsi esimesed kaks peakomponenti riikide ja soolevähi diagnoosi järgi ning nende tihedused, kus on tähistatud AT - Austria, CN - Hiina, DE - Saksamaa, FR - Prantsusmaa, US - Ameerika Ühendriigid.

Jooniselt 2 on näha, et selget klasterdumist riikide vahel ei ole, kuid teistest populatsioonidest eristub mõnevõrra Austria, mille andmepunktid eristuvad keskmisest kõrgemate PC2 väärtuste järgi. Üsna hajutatult asetsevad Hiina, Prantsusmaa ja Ameerika Ühendriikide andmepunktid,

kusjuures Saksamaa andmepunktid katavad pigem väikese ala. Samuti on märgata, et soolevähi ja kontrollrühmad ei erine üksteisest.

4.2. Soolevähi ennustusmudelite loomine

Järgnevalt kirjeldatakse mikrobioomi andmetel põhinevate soolevähi ennustusmudelite loomist. Esmalt tutvustatakse andmete eeltöötlemist, peatükis 3.2. kirjeldatud transformatsioonide rakendamist ning peatükis 3.3. kirjeldatud masinõppe meetodite kasutamist. Andmete töötluseks ja analüüsimiseks kasutatakse rakendustarkvara R (versioon 4.0.2). Peamisteks kasutatavateks pakettideks on *dplyr*, *compositions*, *tidymodels*, *pROC*.

4.2.1. Andmestiku eeltöötlus ja transformatsioonid

Esialgses andmestikus esines kokku 849 bakteri liiki. Suur osa bakteriliikidest on aga harva esinevad, mistõttu filtreeritakse mikrobioomi analüüsimisel sageli harvad liigid andmestikust välja. Kuigi ka mittesagedad liigid võivad omada olulist rolli haiguse prognoosimisel, siis arvutusmahukuse vähendamiseks kasutatakse ka kõnealuses magistritöös alamandmestikku, kus esialgse 849 bakteriliigi asemel on alles 176 bakteriliiki. Filtreerimise aluseks on bakteriliigi levimus üle kõigi proovide ning töös kasutatava alamandmestiku jaoks kaasatakse bakteriliigid, mida detekteeriti vähemalt 50% indiviidide proovides.

Esiteks teisendatakse bakteriliikidele vastavad väärtused suhtelisteks sagedusteks (edaspidi TSS transformatsioon - *total sum scaling*). Selleks jagatakse iga rea väärtus läbi oma rea summaga ning saadakse bakterite suhtelised sagedused. Peatükis 3 kirjeldatud log-suhte teisenduste rakendamiseks asendatakse esmalt kõik nullilised väärtused andmestikus esineva väikseima positiivse väärtusega. Paariviisilise ning tsentreeritud log-suhte transformatsioonide jaoks kasutatakse vastavalt paketi *compositions* funktsioone *pwlr* ja *clr*.

Aditiivse log-suhte transformatsiooni jaoks rakendatakse samast paketis *compositions* defineeritud funktsiooni *alr*. ALR transformatsiooni puhul kasutatakse ühel juhul nimetajas referents-liigina bakteriliiki *Hungatella hathewayi* [ref_mOTU_v2_0882] (ALR1), mille arvukus on Wirbeli jt artikli põhjal soolevähi puhul muutunud, ning teisel juhul liiki *Dorea longicatena* [ref_mOTU_v2_4203] (ALR2), mille arvukus võrrelduna tervete inimestega muutunud ei ole.

Seejärel ühendatakse metaandmete (veergudes isiku unikaalne kood, sugu, vanus, grupitunnus, KMI) ja teisendatud mikrobioomi andmestikud (veergudes isiku unikaalne kood, bakteriliigid)

kasutades ühendamise tingimuseks isikute unikaalseid koode. Lisaks kodeeritakse sootunnusel (*Gender*) „M“ number üheks ja „F“ number kaheks. Sarnaselt kodeeritakse ümber grupitunnus (*Group*): „CTR“ (kontrollrühm) asendatakse numbriga 0 ja „CRC“ (soolevähiga isikute rühm) numbriga 1. Lihtsuse mõttes filtreeritakse välja vaatlused, mille korral KMI väärtus puudub, kokku 8 vaatlust.

4.2.2. Masinõppealgoritmide rakendamine

Mõlema (elastne võrk ja juhumets) masinõppemeetodi jaoks defineeritakse funktsioon, mis võtab argumendiks treenimiseks kasutatava riigi, testimiseks kasutavad riigid ning andmestiku (TSS, CLR, ALR1, ALR2, PWLR teisendustega andmestikud). Funktsiooni üldine skeem on järgmine: esmalt eraldatakse vastava riigi järgi treeningandmestik, mida kasutades treenitakse mudel, ning seejärel rakendatakse mudelit teiste populatsiooni andmetele (testandmed) ning arvutatakse mudeli täpsusenäitajad.

Masinõppemudelite loomiseks kasutatakse paketti *tidymodels*. Järgnevalt kirjeldatakse põhilise samme, mida mudelite treenimiseks rakendatakse. Esmalt defineeritakse mudelite parameetrite hindamiseks kasutatud ristvalideerimise protseduur, kusjuures kasutatakse 5-kordset ristvalideerimist:

```
andmed_cv <- rsample::vfold_cv(treeningandmed, v = 5, repeats = 1)
```

Juhumetsa algoritmi jaoks defineeritakse mudel selliselt, et treenitavateks parameetriteks on *min_n* ja *mtyr*, vastavalt ühe otsustuspuu ehitamisel minimaalne vaatluste arv grupis A_j ($j \in 1, \ldots, k$) ning ühe puu igas sõlmes vaadeldavate juhuslikult valitud tunnuste arv:

```
juhumets_mudel <- rand_forest(
    mtry = tune(),
    trees = 10000,
    min_n = tune()
    ) %>%
    set_mode("classification") %>%
    set_engine("ranger", importance = "permutation")
```

Elastse võrgu korral treenitakse parameetreid *penalty* ja *mixture* ehk vastavalt λ ja α valemis 10:

```
elastne_vork_mudel <- logistic_reg(
    penalty = tune(),
    mixture = tune()
    ) %>% set_engine("glmnet")
```

Seejärel lubatakse parameetritel väärtusi valida 100 juhuslikult valitud parameetrite kombinatsiooni hulgast ning parim mudel valitakse ristvalideerimisvea väärtuse järgi:

```
tuned_model <- defineeritud_mudel %>%
    tune::tune_grid(
        resamples = andmed_cv,
        grid = 100,
        metrics = yardstick::metric_set(yardstick::roc_auc)
    )
```

Siinkohal tuleb mainida, et PWLR-teisenduse ja juhumetsa kombinatsiooni jaoks kasutatakse suure arvutusmahukuse tõttu eelnevas sammus parameetriks `grid=50`.

Valitud parimat mudelit rakendatakse iga ülejäänud riikide järgi filtreeritud testandmestikele ning arvutatakse vastavad AUROC täpsuse näitajad. Selline skeem tehakse läbi kõikide riikide, algoritmide ja teisenduste kombinatsioonide jaoks. Saadud tulemused salvestatakse RDS faili.

```
prognoos <- stats::predict(
    loplik_mudel,
    type = "class",
    new_data = testimine)
train_probs <- predict(
    loplik_mudel,
    type = "prob",
    new_data = testimine
    ) %>%
    bind_cols(obs = testimine$Group) %>%
    bind_cols(prognoos)
roc_auc_value <- roc_auc(train_probs, obs, .pred_0)
```

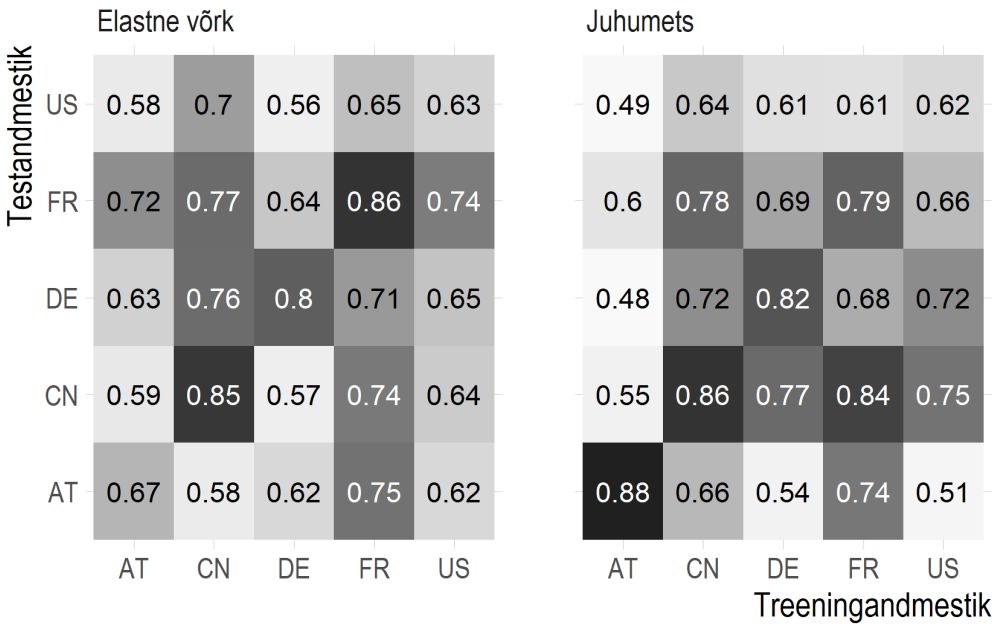
5. Meetodite võrdlus

Analüüs viiakse läbi analoogiliselt artiklis [26] esitatud skeemile, kus iga riigi jaoks treenitakse mudel soolevähi diagnoosimiseks ning saadud mudelit rakendatakse kõikide teiste riikide andmetele. Mudeli prognoosivõime hindamiseks kasutatakse AUROC väärtusi. Ennustusvõime hindamiseks populatsioonis, kus mudel on treenitud, kasutatakse ristvalideerimisviga.

Joonistel kasutatakse iso2 riigikoode: Austria - AT, Hiina - CN, Saksamaa - DE, Prantsusmaa - FR, Ameerika Ühendriigid - US.

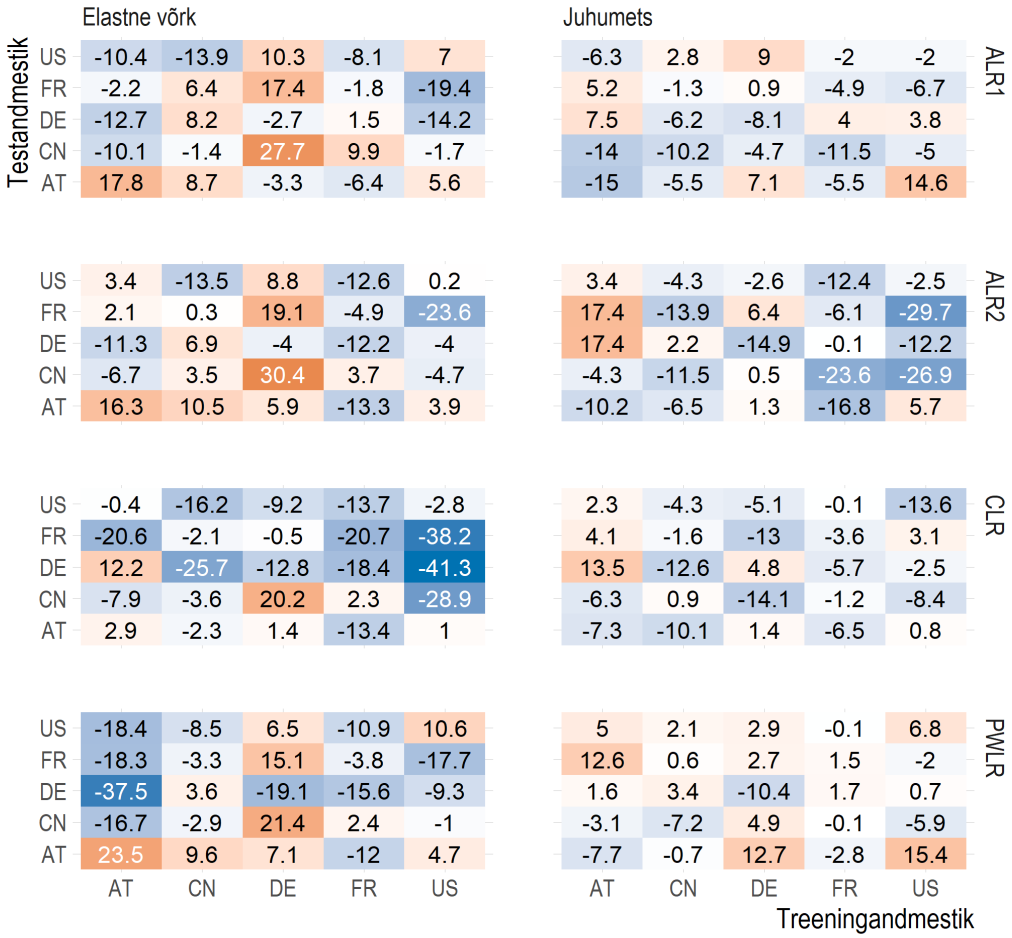
5.1. Mudelite üldistatavus

Enamasti kasutatakse mikrobioomi andmetele prognoosimudelite loomiseks bakterite suhtelisi arvukusi (TSS-transformatsioon). Järgnevalt kirjeldatakse suhtelistele sagedustele ehitatud soolevähi ennustusmudelite prognoosivõimet teistes populatsioonides. Seejärel võrreldakse tulemusi mudelitega, mis kasutavad log-suhte transformatsioone.



Joonis 3. TSS-teisendusega andmestiku AUROC väärtused (kui treening- ja testandmestiku riigid on erinevad) ning ristvalideerimisviga (muul juhtudel) treening- ja testandmestike lõikes elastse võrgu ja juhumetsa korral.

Jooniselt 3 on näha, et kasutades suhteliste sageduste andmestikku annavad erinevate riikide andmetel treenitud mudelid testandmetel väga erinevaid tulemusi mõlema algoritmi puhul. Üldiselt on märgata prognoosivõime kahanemist, sest AUROC väärtused väljaspool diagonaali (alt vasakult üles paremale) on madalamad. Teistest paremini käituvad Prantsusmaa ja Hiina, juhumetsa korral ka Saksamaa andmetel treenitud mudelid. Kõige väiksemad AUROC väärtused on aga mõlema meetodi korral Austria andmetega treenitud mudeli kasutamisel teiste riikide andmetel.

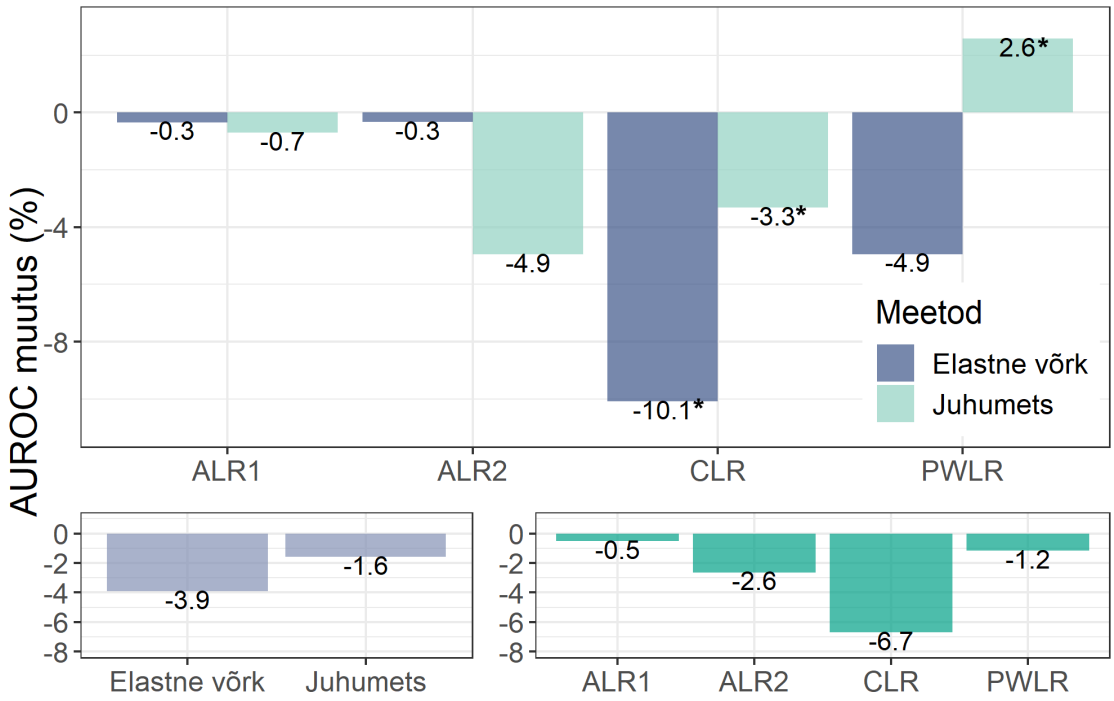


Joonis 4. AUROC ja ristvalideerimisvea keskmised protsentuaalsed muutused vastavatest TSS-teisenduse väärtustest teisenduste (ALR1, ALR2, CLR, PWLR), meetodite (elastne võrk ja juhumets) ning treening- ja testandmestike korral.

Joonisel 4 on esitatud AUROC ja ristvalideerimisvigade protsentuaalsed muutused võrreldes TSS-teisenduse väärtustega kõikvõimalike log-suhte teisenduste, meetodite ning treening-testandmestike lõikes. Joonisel 12 (lisas) on esitatud vastavad AUROC absoluutväärtused. On näha, et tulemused on üsna varieeruvad. Näiteks USA andmetel treenitud elastse võrgu meetodi ja CLR-teisenduse

kombinatsiooniga mudel on Saksamaa andmetele rakendatult lausa 41,3% võrra kehvem kui TSS-teisendusega. Samas Saksamaa andmetega treenitud (elastne võrk ning ALR2-teisendus) mudeli rakendamisel Hiina andmetele on võit 30,4% võrreldes suhteliste sagedustega.

Küll aga on ka selliseid riigi-meetodi-teisenduse kombinatsioonidel treenitud mudeleid, mille testandmetele rakendamine ei parane ega halvene märkimisväärselt võrreldes TSS-teisendusega. Näiteks Prantsusmaa-juhumetsa-CLR kombinatsiooniga treenitud mudeli rakendamisel USA andmetele väheneb AUROC väärtus 0,1% võrra võrreldes TSS-transformatsiooniga. Kokkuvõttes ei eristu selget olukorda, mille korral mõni log-suhte transformatsioon TSS transformatsioonist silmnähtavalt paremini töötaks.

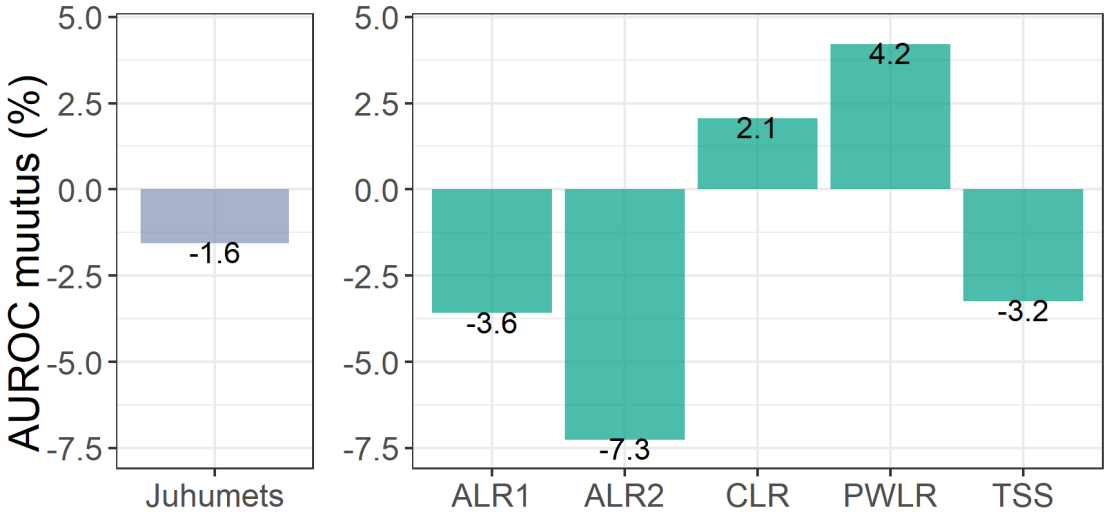


Joonis 5. Log-suhte teisenduste ja masinõppemeetodite keskmised AUROC väärtuste protsentuaalsed muutused võrreldes vastavate TSS-teisenduste väärtustega. * tähistab Wilcoxoni testi põhjal statistiliselt olulist nullist erinevat muutust.

Selleks, et hinnata andmeteisenduste ning kasutatud algoritmide mõju mudelite ennustus- ning üldistatavusvõimele, aggregeeritakse tulemused üle kõigi treening-testandmestik kombinatsioonide. Joonis 5 kirjeldab AUROC väärtuste protsentuaalseid muutusi juhuga, kus on kasutatud TSS transformatsiooni. Selgub, et kõige parema protsentuaalse AUROC muutuse võrreldes TSS-teisenduse AUROC väärtustega teeb juhumetsa ja PWLR-teisenduse kombinatsioon paranedes 2,6% võrra. Teistel juhtudel ei õnnestu tulemusi parandada.

Kõige kehvemini võrreldes suhteliste sagedustega töötab aga elastne võrk ning CLR-teisendus,

mille puhul väheneb AUROC väärtus suisa 10,1% võrra. Kui võrrelda algoritmiti ennustusvõimet log-suhte transformatsioonide rakendamisel, selgub, et keskmised AUROC väärtuste protsentuaalsed muutused nii elastse võrgu kui ka juhumetsa korral ei erine TSS-teisendusest väga palju: väärtused vähenevad vastavalt 3,9% ja 1,6% võrra. Võrreldes keskmisi AUROC väärtuste muutusi transformatsioonide lõikes, siis on näha, et kõige rohkem kaotab võrreldes TSS-teisendusega CLR-teisendus (keskmiselt 6,7% võrra) ning kõige vähem ALR1-teisendus (0,5% võrra).



Joonis 6. Log-suhte teisenduste ja masinõppemeetodite keskmised AUROC väärtuste protsentuaalsed muutused võrreldes vastavate elastse võrgu väärtustega.

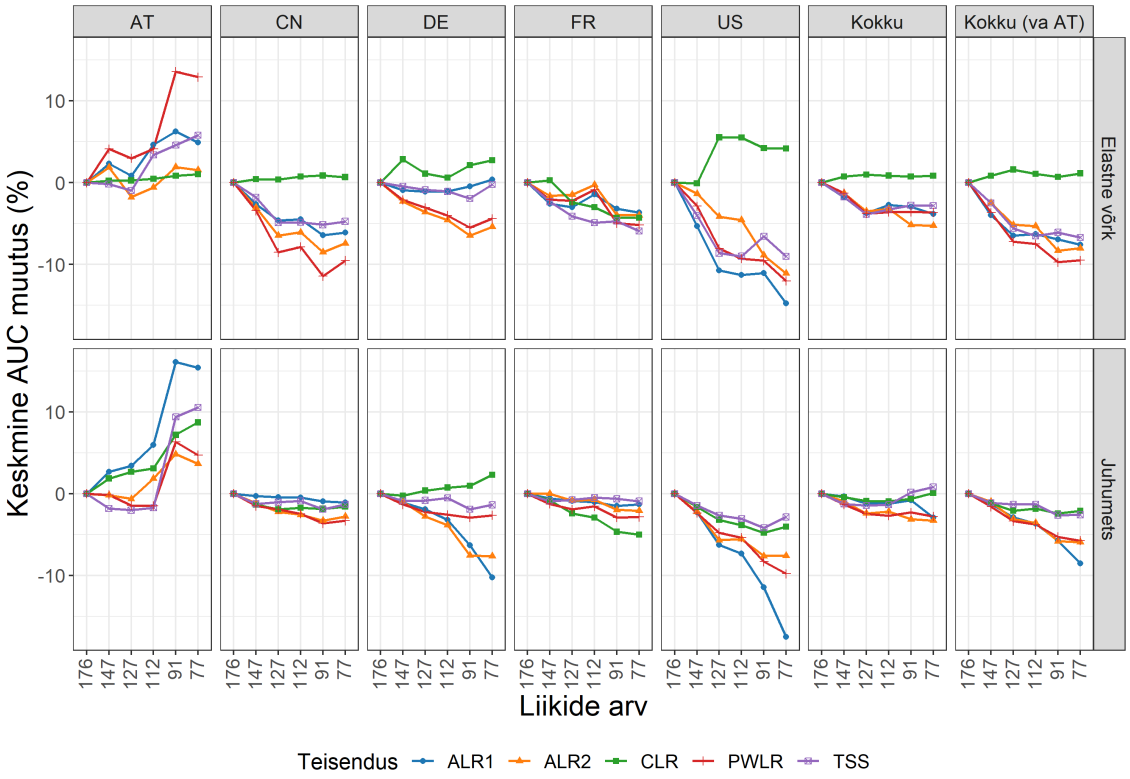
Uurimaks, kumb meetod töötab paremini, leitakse juhumetsa keskmine AUROC väärtuse protsentuaalne erinevus elastse võrgu keskmisest AUROC väärtusest. Jooniselt 6 selgub, et võrreldes elastse võrguga kaotab juhumets 1,6% AUROC väärtusest. Samuti on näha, et võrreldes elastse võrgu meetodiga töötavad CLR- ning PWLR-teisendused juhumetsaga paremini. ALR1, ALR2 ning TSS-teisendused töötavad paremini aga elastse võrgu meetodiga.

5.2. Mudelite üldistatavus alamkompositsioonis

Eelnevalt uuriti juhtu, kus populatsioonis, mille peal mudelit testitakse, on teada mikrobioomi profiil täpselt samade liikide kohta, mis esinesid treeningandmestikus. See on võimalik ainult juhul, kui kõikide populatsioonide andmed on saadud sama bioinformaatilise lähenemise kaudu ning kõik andmed on bioinformaatiliselt samaaegselt analüüsitud. Enamasti aga määratakse mikrobioomi profiil kõikides populatsioonides sõltumatult. Sellisel juhul saadakse populatsioonispetsiifilised andmemaatriksid, mille puhul osad bakteriliigid kattuvad, kuid palju on ka baktereid, mida ühes

populatsioonis tuvastatakse, kuid teises mitte.

Masinõppemeetodite rakendamiseks on aga vajalik, et testandmestikus oleks olemas samad liigid, mida on kasutatud mudeli treenimiseks. Kui populatsioonis, mille peal mudelit treenitakse, on vähem liike kui testpopulatsioonis ning need liigid kattuvad testpopulatsiooni omadega, siis saab testpopulatsiooni andmestikust liike eemaldades kätte analoogse andmetabeli. Kuid enamasti on treeningandmestikus ka liike, mida testandmestikus ei leidu. Sellisel juhul tuleb puuduvad liigid asendada testandmestikus nullide veeruga, mille tõttu võib muutuda masinõppemudelite ennustusvõime.



Joonis 7. Liikide levimuse järgi vähendatud andmestike keskmine AUROC muutus (%) 176 liigiga andmestiku AUROC väärtustest log-suhte teisenduste, masinõppemeetodite ning riikide lõikes.

Järgnevalt uuritakse, kuidas käsitletud mudelite ennustusvõime sõltub sellest, kui testandmestikus ei ole kõik mudelis esinevad liigid tuvastatud. Selleks rakendatakse treeningandmestikul loodud mudeleid testandmestikele, milles on vastavalt 147, 127, 112, 91 ning 77 liiki. Nimetatud arvud lähtuvad sarnaselt alapeatükis 4.3.2 kirjeldatule bakterite levimusele kõikides proovides - vastavale 55%, 55%, 60%, 65%, 70% ning 75%. Analoogiliselt alapeatükis 4.2.1. kirjeldatud viisile rakendatakse vähendatud liikidega andmestikele log-suhte transformatsioone, millele seejärel rakendatakse treenitud mudelid hindamaks mudelite ennustusvõimet. Joonisel 7 on kirjeldatud

AUROC keskmine protsentuaalne muutus esialgsest kõikide liikidega (176 liiki) treenitud mudeli AUROC väärtusest.

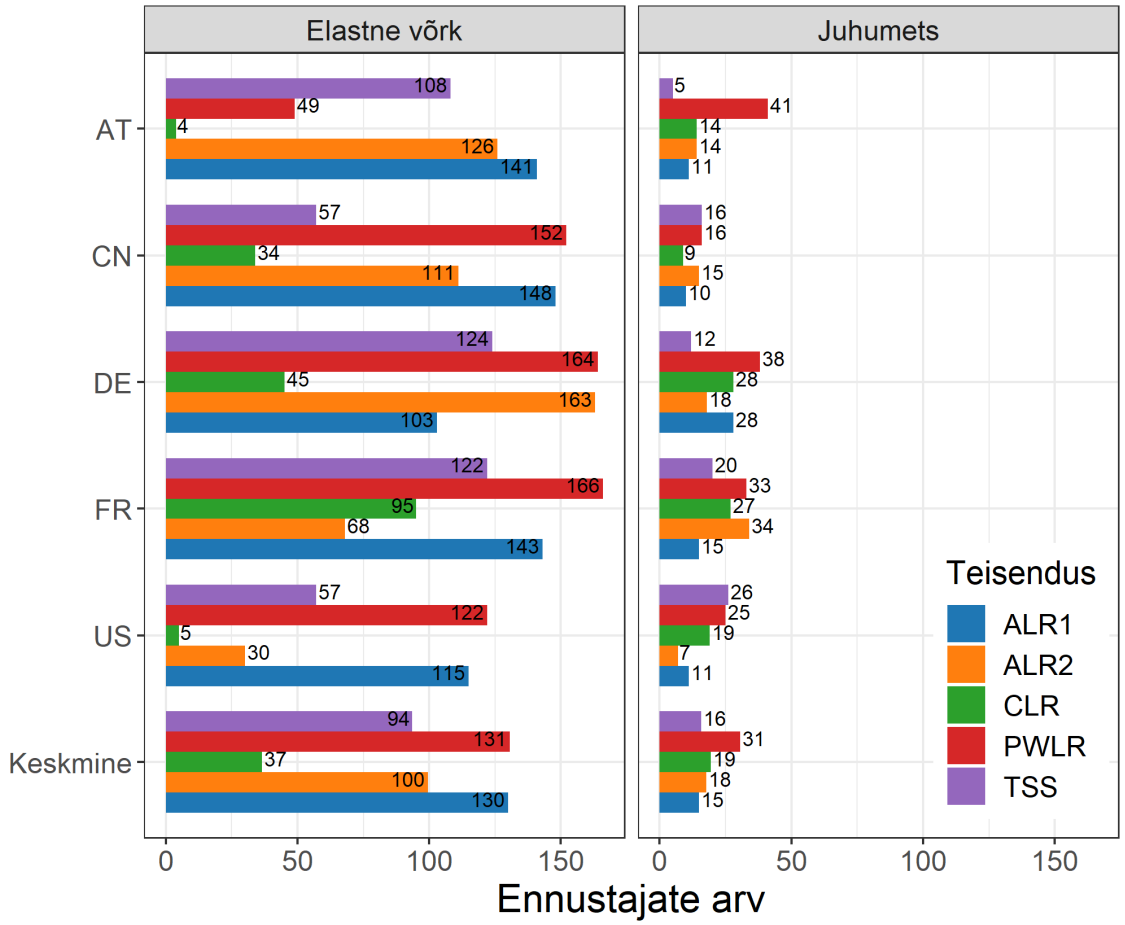
Jooniselt 7 on näha, et kui liike on testandmestikus vähem tuvastatud, siis üldiselt AUROC väärtused mõlema meetodi ning kasutatavate transformatsioonide puhul vähenevad. Näiteks elastse võrgu korral kaotavad AUROC väärtuses kõige rohkem USA ja Hiina andmetel treenitud mudelid peaaaegu kõikide teisenduste korral (va CLR). Juhumetsa korral eristub selgelt ALR1, mille puhul väheneb AUROC väärtus Saksamaa ja USA korral rohkem kui 10% võrra.

Selgub, et liikide vähenemine mõjub kõige paremini Austria andmetel treenitud mudelitele, kus elastse võrgu ja PWLR ning juhumetsa ja ALR1 ning TSS kombinatsioonidega kasvab AUROC väärtus suisa üle 10 protsendi võrra. Peaaegu kõikide riikide (va Prantsusmaa) korral käitub CLR-teisendus elastse võrgu puhul paremini või vähemalt sama hästi kui esialgse (liikide levimus 50%) andmestiku korral. Vaadates tulemusi üle kõikide riikide, on näha, et juhumetsa meetod käitub võrreldes elastse võrguga kõikide teisenduste korral stabiilsemalt.

Kuna Austria andmetel treenitud mudelite ennustusvõime paraneb kõikide teisenduste korral liikide vähenedes, siis on joonisel 7 on esitatud ka keskmised muutused kõikide riikide välja arvatud Austria lõikes. Keskmiselt töötab kõige paremini elastse võrgu korral (nii Austria andmeid arvestades kui ka ilma) CLR-teisendus. Juhumetsa korral aga üle kõigi riigi paraneb AUROC väärtus ainult TSS korral ning kui Austria andmed välja arvata, siis mitte ühegi teisenduse korral.

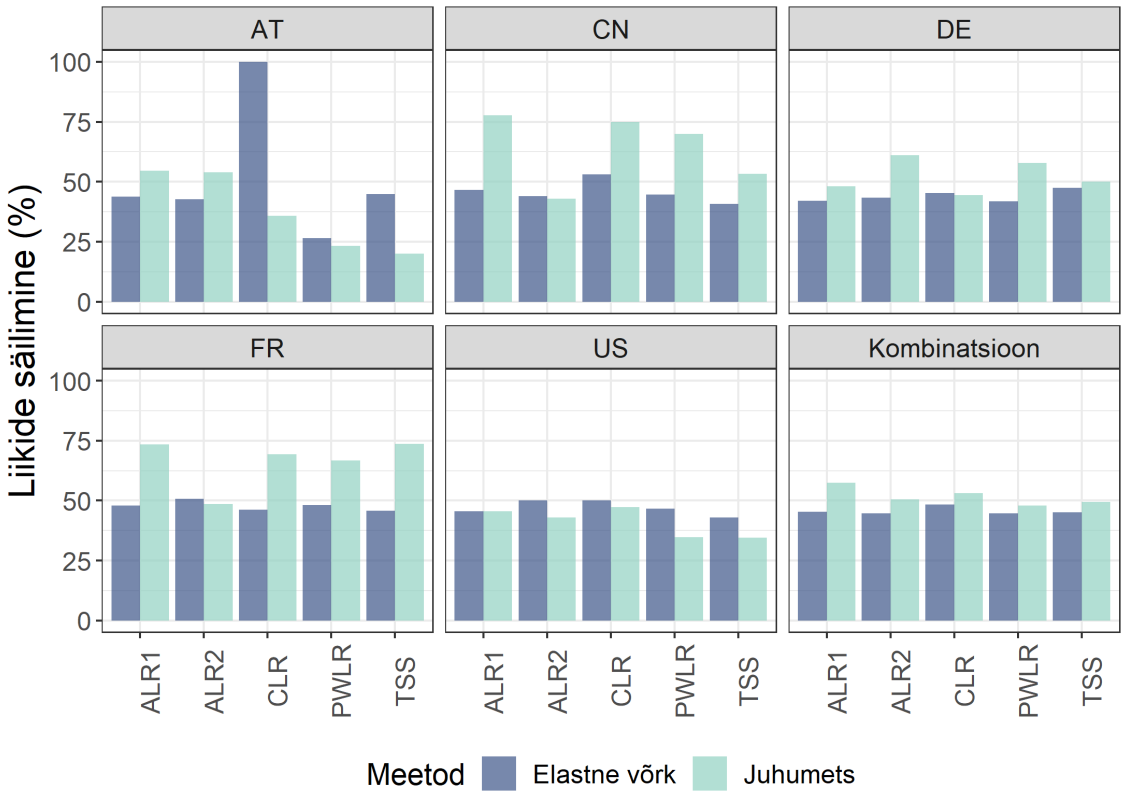
5.3. Ennustajate arv mudelis

Edasi vaadeldakse elastse võrgu, juhumetsa ning erinevate teisenduste ning riikide kombinatsioonidega treenitud mudelites valitud ennustajate arvukusi, esitatud joonisel 8. Arvukused on juhumetsa meetodi puhul subjektiivselt hinnatud kasutades *scree plot'* ning küünarnuki meetodit (lisas joonis 13).



Joonis 8. Mudelitesse valitud ennustajate arvud elastse võrgu ja juhumetsa ning kõikide log-suhte teisenduste ning populatsioonide korral.

Jooniselt 7 on näha, et CLR-teisenduse korral (vähendades liikide arvu) keskmine AUROC väärtus väheneb vaid Prantsusmaa puhul. Seda arvesse võttes ning uurides, kui palju ennustajaid mudel valib (joonisel 8), on näha, et elastse võrgu ja CLR korral on Prantsusmaa puhul mudelisse valitud kõige rohkem ennustajaid. Analoogselt käitub PWLR-teisendus, mille korral on mudelisse valitud ennustajaid kõige vähem Austria puhul ning jooniselt 7 selgub, et just Austria korral saab PWLR-teisendus kõige paremini liikide vähenemisega hakkama.



Joonis 9. Mudelisse (77 liigiga andmestikuga treenitud) valitud ennustajate säilimine protsentides võrreldes esialgse mudelisse valitud ennustajate arvuga log-suhte teisenduste, masinõppemeetodite ja riikide lõikes.

Joonisel 9 on esitatud mudelisse jäävate ennustajate (kui vähendatud andmestikus on 77 liiki) protsent esialgse andmestikuga treenitud mudelisse jäävatest ennustajatest. On näha, et juhumetsa puhul säilib üle poolte juhtudest rohkem ennustajaid, erandid on Austria puhul CLR, PWLR ja TSS teisendused; Hiina puhul ALR2-teisendus; Saksamaa korral CLR-teisendus, Prantsusmaa korral ALR2-teisendus, USA korral aga kõik teisenduses va ALR1. Vaadates üle riikide keskmisi ennustajate säilimiste protsente, võib arvata, et juhumets säilitab kõikide teisendustega elastsest võrgust rohkem esialgses mudelis olevaid ennustajaid ning võib seetõttu olla stabiilsem (võrreldes elastse võrguga).

Arutelu

Peatükis 5.1. selgus, et üleüldiselt on erinevate meetodite-teisenduste, sealjuures TSS-teisenduse korral erinevatel populatsioonidel treenitud ja testitud mudelite AUROC väärtused küllaltki madalad, mõnel juhul jäädes isegi alla 0,5. See võib olla tingitud sellest, et kasutatavad soolestiku mikrobioomi andmed on populatsiooniti väga varieeruva kooslusega, proovid on eri meetoditega kogutud ning erinevalt hoiustatud/säilitatud. Seega tundub, et mikrobioomi abil soolevähki ennustades tasub mudelite ehitamisel keskenduda ühe populatsiooni andmetele, kus proovide kogumise tingimused on homogeensed.

Praktikas on väga oluliseks aspektiks treenitud mudelite üldistatavus mikrobioomi andmetele, milles esineb vähem bakteriliike. Selles töös nähti, et sellisel juhul käitub stabiilsemalt juhumetsa meetod, kuid mõlema masinõppemeetodi korral üldistatavus pigem väheneb. Küll aga eristub teistest populatsioonidest Austria, mille puhul tundub liikide vähenedes üldistatavus kasvavat. Liikide vähenedes väheneb reeglina ka mudelisse jäävate ennustajate arv (va Austria ja elastse võrgu ning CLR-teisenduse kombinatsioonil treenitud mudel).

Kirjanduses on näidatud, et juhumetsa meetodiga töötab väga hästi PWLR-teisendus (peatükk 3.4.). Ka selles töös leiti, et juhumetsa ning PWLR-teisenduse kombinatsioon töötab võrreldes elastse võrgu ja PWLR-teisenduse kombinatsiooniga paremini. Samuti nähtus, et PWLR-teisendus töötas juhumetsa puhul analoogiliselt kirjanduses välja tooduga paremini kui TSS-teisendus. Küll aga on kirjanduses esitatud, et võrreldes TSS-teisenduse ja lassoregressiooni ning CLR-teisenduse ja lassoregressiooni kombinatsioone, töötab viimane kombinatsioon paremini. Selles töös saadi aga vastupidine tulemus, mis võib olla tingitud erinevate objektide vaatlemisest (kirjanduses ei uuritud inimese mikrobioomi) või üsna väikesest valimimahu suurusest.

Töös vaadeldi ALR-teisendust kahel juhul: referents-elemendiks valiti selline bakteriliik, mille arvukus on haigetel ja tervetel muutunud (ALR1); ning bakteriliik, mille arvukus ei ole muutunud (ALR2). Selgus, et TSS ja elastse võrgu kombinatsiooni ning mõlema ALR-teisenduse vahel kuigi suurt erinevust ei ole. Juhumets ja TSS-teisenduse kombinatsioon töötab ALR2-teisendusega võrreldes paremini ja ALR1-teisendusega võrreldes kehvemini. Kirjanduses on spekuleeritud ka selle üle, et kui valida ALR referents-element sobivalt, siis on võimalik tulemusi parandada. Sobiva referent-elemendi korral peavad olema täidetud järgmised tingimused: log-suhte geomeetriad esialgsel kompositsioonil ning ALR-teisendusega peavad olema sarnased, referents-element peaks olema võimalikult vähe varieeruv (peaaegu konstantne) ning populatsioonis levinud. Sel-

liste kriteeriumite järgi saadud tulemused on väga hästi interpreteeritavad, sest jagatakse peaaegu konstantse tunnusega [18]. Et ALR transformatsioon pakkus ennustustäpsuselt võrreldavaid tulemusi TSS-teisendusega, siis tuleks edasi uurida, kuidas mõjutab optimaalse referents-elemendi valik mudelite ennustusvõimet. Kuna ALR transformatsioon rahuldab erinevalt TSS-teisendusest kompositsionaalsete andmete analüüsimise printsiipe ning on täpsemini interpreteeritav, on ALR transformatsiooni kasutamisel olulisi eeliseid.

Vaatamata sellele, et PWLR-teisendust (juhumetsaga) kasutades saadakse häid tulemusi, on selle teisenduse puhul tulemuste interpreteerimine keeruline. Veel enam on PWLR-teisenduse läbiviimine arvutusmahukas, sest bakteriliikide vaheliste kõikvõimalike log-suhete arv kasvab. Ka selles töös puututi sellise probleemiga kokku – kaasates mudelisse terve andmestiku pealt arvutatud paariviisilised log-suhted, jäi Geenivaramu serverites mälumahtu juhumetsa mudelite ehitamiseks väheseks. Seetõttu on PWLR teisenduse kasutamine praktikas raskendatud.

Senini pole kirjanduses suurt tähelepanu saanud mikrobioomi andmetes esinevate nullidega tegelemine. Enamlevinud lähenemine on kõikide nullide korraga väikese mittenegatiivse arvuga asendamine, mis ei pruugi anda kõige paremaid tulemusi. Tulevikus tasub analüüsida erinevate nullidega tegelemise meetodite mõju masinõppemudelite ennustusvõimele.

Kokkuvõte

Magistritöö eemärgiks oli uurida erinevate mikrobioomi andmete eeltöötlemismeetodite, täpsemalt log-suhte transformatsioonide mõju soolevähi diagnoosimiseks kasutatavate masinõppealgoritmide ennustus- ja üldistusvõimele. Töö esimeses osas tutvustati lühidalt inimese mikrobioomi ning selle seost soolevähiga. Teises osas kirjeldati aga üldiselt mikrobioomi andmete tekkeviisi ning põhilisi omadusi.

Töö kolmandas osas anti ülevaade kasutatavast metoodikast. Kirjeldati mikrobioomi andmetele omast kompositsionaalsust ning selle eiramisest tekkivaid probleeme klassikaliste andmeanalüüsi meetoditega, nt korrelatsioonianalüüsiga. Lisaks tutvustati kompositsionaalsete andmete põhiprintsiipe ning log-suhte teisendusi, mille abil on võimalik mikrobioomi andmetele standardseid statistilisi meetodeid rakendada. Seejärel anti ülevaade kahest mikrobioomi andmetega enim kasutatavast masinõppemeetodist: elastne võrk ning juhumets.

Neljandas osas kirjeldati töös kasutatavaid mikrobioomi andmeid, neile log-suhte teistenduste rakendamist ning masinõppemeetodite läbi viimist. Viiendas osas anti ülevaade saadud tulemustest, millele järgnes diskussioon.

Edaspidi on plaanis saadud tulemusi võrrelda spetsiaalselt mikrobioomi andmetele välja töötatud meetodite tulemustega. Lisaks on tulevikus plaanis tulemusi rakendada Eesti soolevähi sõeluuringust kogutud mikrobioomi andmetele. Veel võiks huvi pakkuda log-suhte teisenduste ja masinõppemeetodite kasutamine teistsugustel struktuuridel, näiteks mikrobioomi andmete peakomponentanalüüsi peakomponentidel või erinevatel latentsetel tunnustel.

Kasutatud kirjandus

[1] Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V, Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*. **24**, 392-400.

[2] Krigul, K. L., Aasmets, O., Lüll, K., Org, T., Org, E. (2021). Using fecal immunochemical tubes for the analysis of the gut microbiome has the potential to improve colorectal cancer screening.

[3] Eesti Haigekassa. Jämesoolevähi sõeluuring. URL: `https://www.haigekassa.ee/inimesele/haiguste-ennetus/jamesoolevahi-soeluuring` (vaadatud 2.05.2021).

[4] Wilkinson, J. E., Franzosa, E. A., Everett, C., Li, C. (2021). A framework for microbiome science in public health. *Nature Medicine*.

[5] Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I. (2021). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology*. 12:635781.

[6] Pasolli, E., Truong, D. T., Malik, F., Waldron, L., Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomoc Datasets: Tools and Biological Insights. *PLoS Comput Biol*. **12**(7).

[7] Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Turukalo, T. L., jt. (2021). Application of Machine Learning in Human Microbiome Studies: A review of Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Frontiers in Microbiology*.

[8] Xia, Y., Sun, J., Chen, D-G. (2018). *Statistical Analysis of Microbiome Data with R*. Springer.

[9] Pinto, J. R. (2018). *Statistical methods for the analysis of microbiome compositional data in HIV studies*. IrsiCaixa AIDS researh institute, UVic-UCC.

[10] Gloor, Gregory. B., Macklaim, J. M., Pawlowsky-Glahn, V., Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*. **8**.

[11] Silverman, J. D., Roche, K., Mukherjee, S., David, L.A. (2020). Naught all zero in sequence count data are the same. *Computational and Structural Biotechnology Journal*. **18**, 2789-2798.

[12] Coyte, K. Z., Rakoff-Nahoum, R. (2019). Understanding Competition and Cooperation within the Mammalian Gut Microbiome. *Current Biology*. **29**, R538-R544.

[13] Gordon-Rodriguez, E., Quinn, T. P., Cunningham, J. P. (2021). Learning Sparse Log-Ratios for High-Throughput Sequencing Data.

[14] Tolosano-Delgado, R., Talebi, H., Khodadadzadeh, M., van den Boogaart, K. G. (2019). On machine learning algorithms and compositional data. *Conference paper, CoDaWork2019*. 172-175.

[15] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. (2011). *Lecture Notes on Compositional Data Analysis*.

[16] Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.(2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*. **35**, 279-300.

[17] Greenacre, M. (2019). Amalgamations are valid in compositinal data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Applied Computing and Geosciences*. **5**(2020).

[18] Greenacre, M., Martínez-Álvaro, M., Blasco, A. (2021). Compositional data analysis of microbiome and any-omis datasets: a revalidation of the additive logratio transformation. *bioRχiv*

[19] Tolosana-Delgado, R. Pairwise log ratio transform. URL: `https://rdrr.io/cran/compositions/man/pwlr.html`. (vaadatud 24.04.2021).

[20] James, G., Witten, D., Tibshirani, R., Hasie, T. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

[21] Pereira, J. M., Basto, M., Ferreira da Silva, A. (2015). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*. **39** (2016), 634-641.

[22] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. **33**, 1-22.

[23] Hastie, T., Tibshirani, R., Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*. Springer.

[24] Zhang, M., Shi, Wenjiao. (2019). Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences*

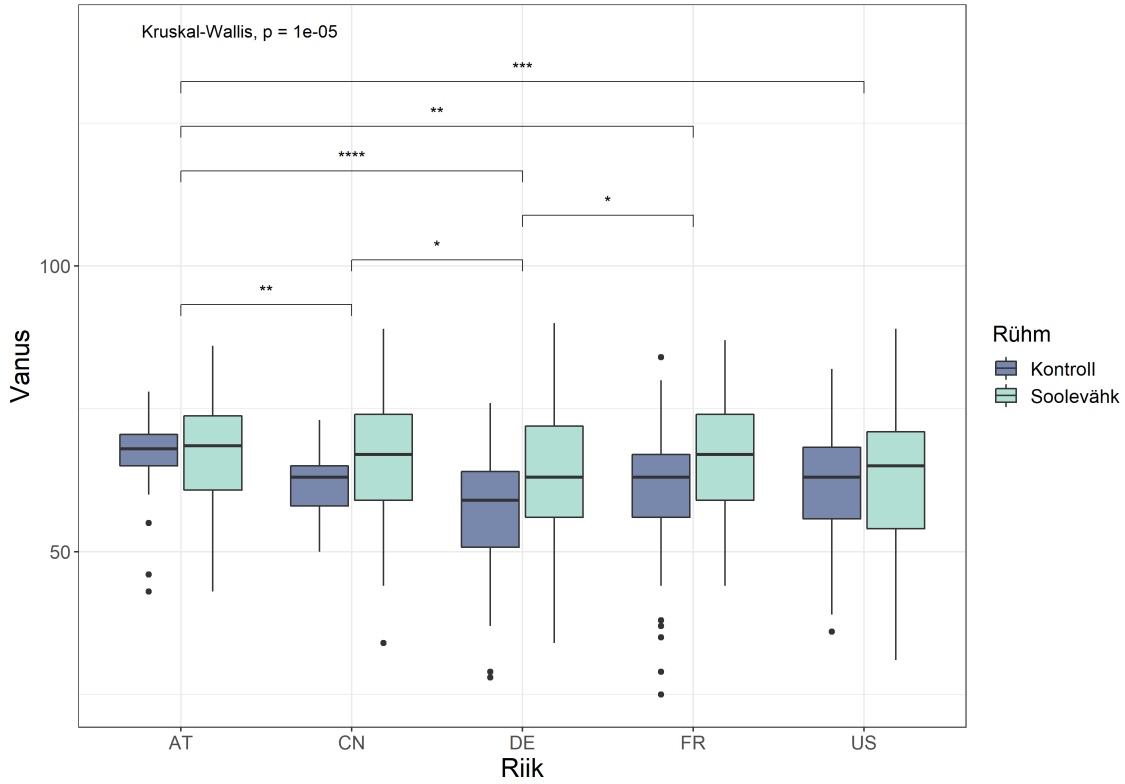
[25] Quinn, T. P., Erb, I. (2020). Interpretable Log Contrasts for the Classification of Health Biomarkers: a New Approach to Balance Selection. *American Society for Microbiology Journals*.

[26] Wirbel, J., Pyl, T. P., Kartal, E., Zych, K., jt. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*. **25**, 679-689.

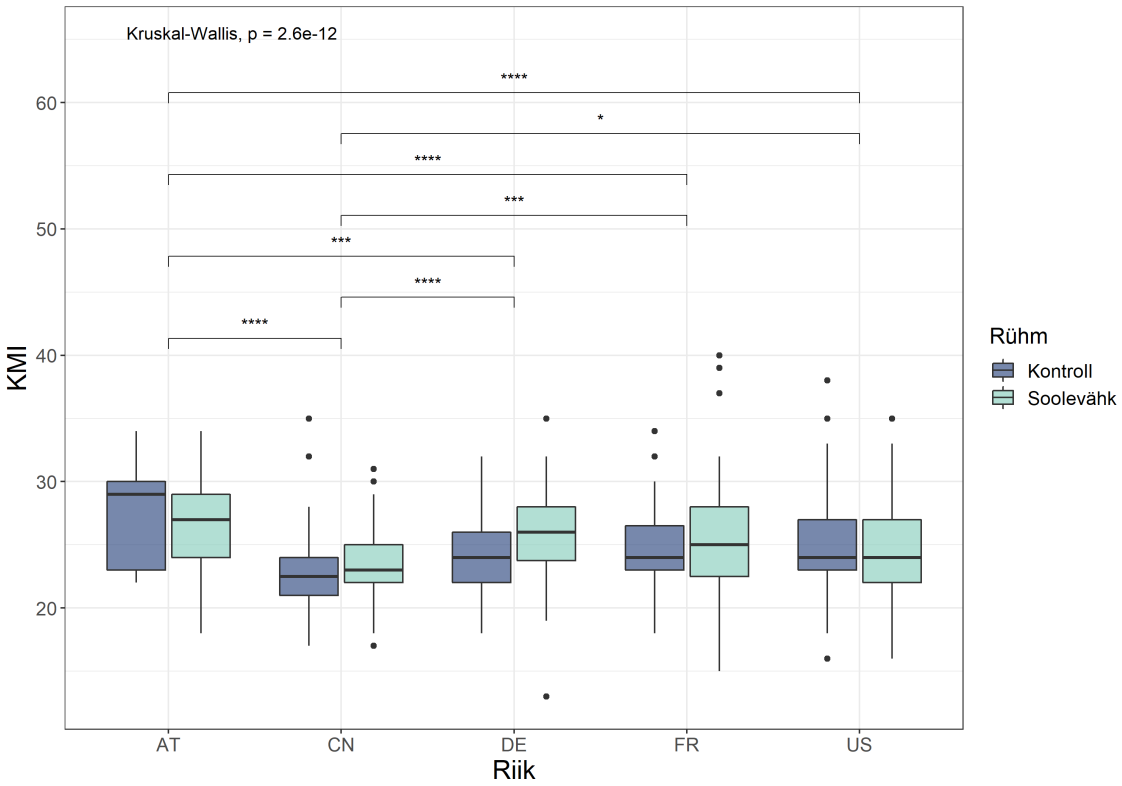
[27] Gloor, G. B. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*. **62**(8).

Lisad

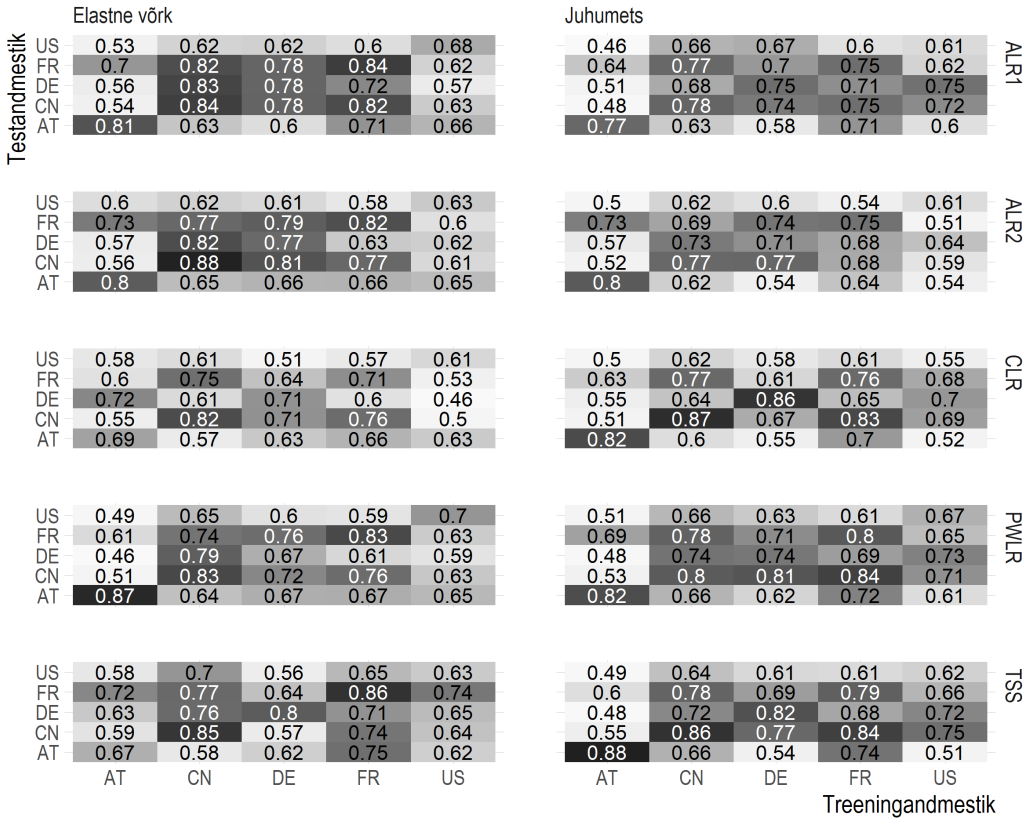
Joonised



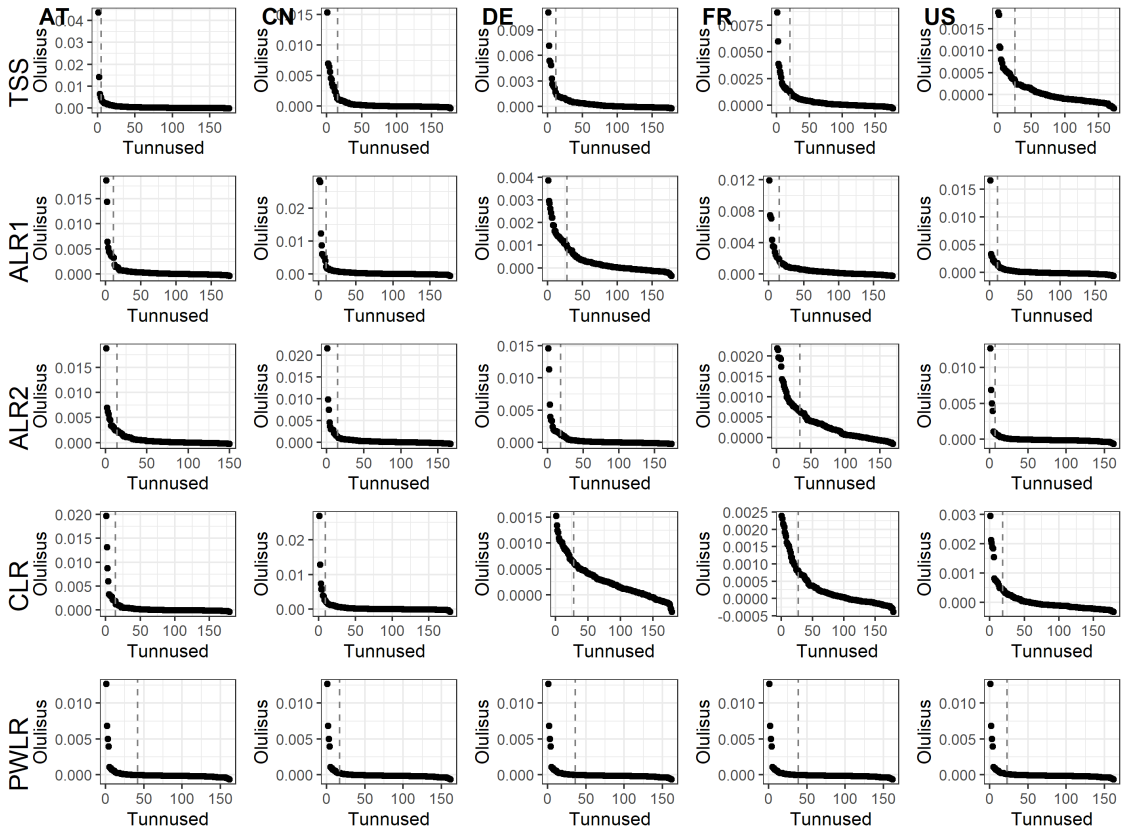
Joonis 10. Analüüsis kasutatava andmestiku indiviidide vanuseline jaotus kontroll- ja soolevähirühmades.



Joonis 11. Analüüsis kasutatava andmestiku indiviidide kehamassiindeksi jaotus kontroll- ja soolevähirühmades.



Joonis 12. Log-suhte teisenduste ja masinõppemeetodite kombinatsioonide AUROC väärtused riikide lõikes, kus peadiagonaalidel (blokkides alt vasakult üles paremale) on ristvalideerimisvead ning väljaspool diagonaale AUROC väärtused.



Joonis 13. Juhumetsa mudeli ennustajate olulisus log-suhte teisenduste ning treeningpopulatsioonide lõikes.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristina Muhu,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Log-suhte teisendustel põhinevad ennustusmudelid soolevähi diagnoosimiseks mikrobioomi andmetelt“, mille juhendaja on Oliver Aasmets, 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.

Kristina Muhu
25.05.2021