TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Eerik Sven Puudist

Kaalulangetuse edukust ennustavate iseloomujoonte määramine

Magistritöö (15 EAP)

Juhendajad: Uku Vainik, PhD
Raivo Kolde, PhD
Kadri Arumäe, PhD

Tartu 2024

Kaalulangetuse edukust ennustavate iseloomujoonte määramine

Lühikokkuvõte:
Käesolev uurimus vaatleb isiksuseomaduste seoseid kehamassiindeksi (KMI) muutustega kasutades varasemate uuringutega võrreldes detailsemat isiksuse mõõtmise küsimustikku ning suuremat valimit ($N > 45000$). Kasutades regressioonmudeleid (lineaarsed mudelid, juhuslik mets, XGBoost), Gaussi segumudeleid ning faktoranalüüsi uurime isiksuseomaduste seoseid inimese maksimaalse KMIga, KMI langusega ning KMI taastõusuga. Eesmärgiks on ära kaardistada nende sündmustega seotud iseloomu jooned, et aidata kaasa isikupärastatud kaalulangetusravi välja töötamisele. Toome välja, et KMI on seotud peaaegu kõikide iseloomu aspektidega, kusjuures seosed on loomult lineaarsed. KMI muutust ennustavad tegurid on aga sugude lõikes erinevad ning sestap on oluline kaalulangetuse ravi metoodika koostamisel sugupooli eraldi käsitleda.

Võtmesõnad:
isiksus, suur viisik, KMI, soolised erinevused, lineaarsed mudelid, faktoranalüüs, Gaussi segumudelid

CERCS:
P160: Statistika, operatsioonanalüüs, programmeerimine,
finants- ja kindlustusmatemaatika
P176: Tehisintellekt
S260: Psühholoogia

Identification of Character Traits Associated with Successful Weight Loss

Abstract:
This study examines the relationship of personality traits with changes in body mass index (BMI) using a more detailed personality measurement questionnaire and a larger sample ($N > 45000$) compared to previous studies. Using regression models (linear models, random forest, XGBoost), Gaussian mixture models, and factor analysis, we examine the relationships of personality traits with a person's largest historical BMI, largest BMI decline, and BMI regain. The goal is to find which character traits influence those outcomes the most to assist in developing personalized weight loss treatment plans. We point out that BMI is related to almost all aspects of one's personality. Those relationships seem to be linear. However, the factors predicting the change in BMI are different between genders, and therefore both genders should be addressed separately in the development of weight loss methodologies.

Keywords:
personality, big five, BMI, gender differences, linear models, factor analysis, Gaussian mixture models

CERCS:
P160: Statistics, operation research, programming, actuarial mathematics
P176: Artificial intelligence
S260: Psychology

Kaalulangetuse edukust ennustava iseloomujoonte määramine

Andmeteadus (MSc)
#UniTartuCS
2024

Autor:
Eerik Sven Puudist
Juhendajad:
Uku Vainik, PhD
Raivo Kolde, PhD
Kadri Arumäe, PhD

TARTU ÜLIKOOL
arvutiteaduse instituut

ikoonide allikas: flaticon.com

Identification of Character Traits Associated with Successful Weight Loss

Data Science (MSc)
#UniTartuCS
2024

Author:
Eerik Sven Puudist
Supervisors:
Uku Vainik, PhD
Raivo Kolde, PhD
Kadri Arumäe, PhD

UNIVERSITY OF TARTU
Institute of Computer Science

icons taken from: flaticon.com

Sisukord

1 Sissejuhatus

Ülekaal ning sellest johtuvad probleemid muutuvad arenenud maailmas aina aktuaalsemaks. Liigne kehakaal ei mõjuta negatiivselt mitte ainult inimese füüsilist tervist, vaid tekitab ka psühholoogilisi ja sotsiaalseid probleeme. Efektiivsemad meetmed ülekaalu ennetamiseks ja ravimiseks on seetõttu olulised nii üksikisiku heaolu edendamiseks kui ka tervishoiu süsteemile langeva koormuse vähendamiseks ning ühiskonna kui terviku hea käekäigu tagamiseks.

Paraku on ülekaalust vabanemine aga äärmiselt keeruline protsess. Kuna tõusma hakanud kehakaalu taga on eri inimestel erinevad füsioloogilised, psühholoogilised ning elustiilist tulenevad põhjused, ei ole kõikide patsientide puhul võimalik rakendada samu ravi metoodikaid.

Personaliseeritud raviplaani koostamisel on tarvis arvesse võtta patsiendi tervise andmete kõrval ka tema iseloomu eripärasid. See eeldab konkreetselt kehakaaluga seotud isiksuse joonte hindamiseks loodud mõõtevahendi loomist, mille esimeseks saamuks on antud kontekstis oluliste isiksuse joonte kaardistamine.

Käesolaves uurimuses vaatleme isiksuse seoseid kolme suurema väljundiga: uurime, millised iseloomujooned aitavad ette ennustada ülekaalu teket, kaalu languse edukust ning kaalu taas tõusma hakkamist peale edukat langetust. Kasutades ennustavaid mudeleid, uurime millisel määral iseloom üldse antud tulemeid ennustada aitab. Seejärel kasutame faktoranalüüsi, et oluliseimad psühholoogilised konstruktid üles leida.

Eesti geenidoonarite hulgas on seni läbi viidud kaks isiksusetesti: ajavahemikus 2009 – 2016 vastasid umbes 3500 doonarit NEO-PI 4 testile, aastatel 2021 – 2022 täitsid 77000 doonarit 100NP küsimustiku. Lisades 100NP küsimustiku vastustele andmed inimeste kehamassiindeksi (KMI) muutuse kohta on võimalik uurida isiksuse ja KMI seoseid kasutades varasemate uuringutega võrreldes detailsemaid isiksuseomaduste kaardistusi suurel ja mitmekesisel valimil.

Loodame, et antud uurimus annab oma panuse senisest efektiivsemate kaalulangetuse metoodikate välja töötamisele ning et osad kehakaalu muutusi uurides leitud seaduspärad on üldistatavad ka teistele positiivsetele pingutust nõudvatele elumuutustele.

E: ekstravertsus	A: sotsiaalsus	C: meelekindlus	N: neurootilisus	O: avatus
soojus	usaldus	asjatundlikkus	ärevus	avatus fantaasiale
seltsivus	siirus	korralikkus	vaenulikkus	avatus kunstile
kehtestavus	omakasupüüdmatus	kohusetundlikkus	masendus	avatus tunnetele
aktiivsus	järeleandlikkus	eesmärgipärasus	enesekontroll	avatus teguviisidele
positiivsed emotsioonid	tagasihoidlikkus	enesedistsipliin	impulsiivsus	avatus mõtetele
elamustejanu	osavõtlikkus	kaalutlemine	abitus	avatus väärtustele

Tabel 1. Suure viisiku 30 alamomadust

2 Ülevaade isiksus kirjeldavatest teooriatest ja mõõtevahenditest

Isiksust kirjeldavate matemaatiliste mudelite ajalugu ulatub tagasi möödunud sajandi neljakümnendatesse aastatesse mil psühholoogid hakkasid kaardistama inimeste kirjeldamiseks kasutatavaid loomuliku keele sõnu eeldades, et olulisemate inimeste erinevusi kirjeldavate omaduste kohta on välja kujunenud keelelised mõisted. Inglise keelele keskenduvad Allport ja Odbert tuvastasid peaaegu 18000 sellist sõna, mille nad jagasid neljaks suuremaks rühmaks: iseloomuomadused (e.g. agressiivne, kartlik), meeleseisundid (e.g. rõõmus, hirmul), hinnangud (e.g. suurepärane, vääriline) ning inimese kehaga seotud mõisted (John et al., 2008). Keskendudes edaspidi eelkõige iseloomuomadusi kirjeldavatele mõistetele hakkasid psühholoogid faktoranalüüsi abil tuvastama peamiseid isiksuse mõõtmeid. Alates möödunud sajandi üheksakümnendate aastate keskpaigast on levinuimaks isiksuseomaduste mudeliks kujunenud suure viisiku mudel (John et al., 2008, joonis 4.1). Samas ei hõlma suur viisik sugugi mitte kõiki inimese iseloomu aspekte, mistõttu on uuematesse mõõtevahenditesse lisada ka täiendavaid komponente.

2.1 Suure viisiku mudel

Suure viisiku mudel kirjeldab iseloomu viie peamise mõõtme kaudu, millest igaühe võib omakorda jagada kitsamateks ja veel kitsamateks omadusteks. Nende viie faktori nimetamiseks on välja pakutud hulganiselt erinevaid märksõnu, kuid levinumateks mõisteteks on kujunenud ekstravertsus (ingl. *extraversion*), sotsiaalsus (ingl. *agreeableness*), meelekindlus (ingl. *conscientiousness*), neurootilisus (ingl. *neuroticism*) ja avatus (ingl. *openness*, kasutatakse ka ‘avatus kogemusele’), mille esitähtedest tuleb ingliskeeles kokku lühend *OCEAN* (Kanger, 2013; Kangur, 2012; John et al., 2008, joonis 4.2).

Igaühe neist viiest dimensioonist saab omakorda jagada näiteks NEO PI mudeli järgi kuueks alamskaalaks. Need kokku 30 isiksuse tahku on ära toodud tabelis 1.

Need viis dimensiooni paistavad olevat suhteliselt kultuuride ülesed, vähemasti moodsa elukorraldusega ühiskondades[1]. Kõige muutlikum paistav olevat viies diemnsioon,

[1]Boliivia põliselaanike hulga läbi viidud uurimuses suudeti tuvastada vaid kaks isiksuse dimensiooni (Gurven et al., 2013). See annab alust oletada, et viie dimensiooni mudel ei pruugi kehtida väljaspool

avatud kogemusele, mis teatud mõõtevahendite korral korral kajastab ka mässumeelsust ning kehtivatest konvensioonidest irdumist (John et al., 2008).

Antud viiest dimensioonist on räägitud eelkõige täiskasvanute puhul, kuid vähemasti ekstravertsuse, sotsiaalsuse ning meelekindluse dimensioonid kujunevad välja juba kuuendaks eluaastaks. Teismeeas saab peale viie peamise dimensiooni eristada ka ärritatavuse, aktiivsuse ja teistest sõltuvuse dimensioone (John et al., 2008).

2.2 Elusündmuste ennustamine isiksuseomaduste põhjal

Isiksuseomadused mõjutavad, kuidas inimene erisuguseid nähtuseid enda jaoks mõtestab, kuivõrd ta eri stiimulitele (e.g. arsti soovitatud raviplaan) reageerib ning kuidas ta enda sotsiaalset ja ainelist keskkonda valib ja ümber kujundab (John et al., 2008). See annab alust oletada, et isiksuseomadused aitavad ette ennustada erisuguseid sündmuseid inimese elus.

Peaaegu kõik suure viisiku domeenid aitavad ette ennustada inimese füüsilisi tervise näitajaid. Mitmed uuringud on kinnitatud meelekindluse seost tervislikuma eluviisi ning pikaealisusega. Meelekindlamad inimesed hoiduvad suurema tõenäosusega suitsetamisest ning hoiavad oma söömis- ja liikumisharjumused korras. Madal sotsiaalsus aitab ennustada südame-veresoonkonna haiguseid, kõrge neurootilisus muudab haigustega toimetuleku keerulisemaks, kõrge ekstravertsus aga lihtsamaks, kuna aitab inimesel kogeda rohkem sotsiaalset toetust (John et al., 2008).

Noorukieas ennustavad madal sotsiaalsus ja meelekindlus kuritegevust, samas kui kõrge neurootilisus ja madal meelekindlus seostuvad ärevuse ja depressiooniga (John et al., 2008).

Meelekindlus ja avatus ennustavad kõrgemat akadeemilist suutlikust ning paremaid tulemusi enamikul ametikohtadel. Ülejäänud suure viisiku mõõtmed muutuvad tähtaks konkreetsete ametite kontekstis: sotsiaalsus ja neurootilisus ennustavad meeskonnatöö võimekust, ekstravertsus edukust müügis ja juhtivatel kohtadel. Samuti suurendab kõrge neurootilisus läbipõlemise ja töökohtade sageda vahetamise tõenäosust (John et al., 2008).

Teismeeas ennustab madal meelekindlus ja ekstravertsus ning kõrge neurootilisus probleeme suhetes vanematega, madal sotsiaalsus ja ekstravertsus aga välja jäätust omaealiste gruppidest (John et al., 2008).

Kõrge ekstravertsus ja meelekindlus ning madal neurootilisus viivad kauakestvate ja õnnelike suheteni (John et al., 2008).

---

kaasaegset kultuuriruumi.

Samas on aga kõiki viite isiksuse mõõdet suudetud tuvastada ka inimesest erinevate loomaliikide puhul: inimahvide, kasside, koerte, eeslite, sigade ja kaheksajalgade hulgas on näidatud individuaalseid erisusi ekstravertsuse, neurootilisuse ning sotsiaalsuse osas, loomadel on kirjeldatud individuaalseid erinevusi ka avatuse (e.g. uudishimu ja mängulisuse) osas. Šimpansite puhul on kirjeldatud erinevusi ka meelekindluse osas (John et al., 2008).

Kõrge ekstravertsus ja meeste puhul ka madal neurootilisus ennustavad kõrgemat sotsiaalset staatust (John et al., 2008).

2.3 Isiksuseomaduste mõõtevahendid

Suure viisiku 30 alaskaala mõõtmise kullastandardiks on pikka aega peetud NEO-PI-R küsimustikku, mida on aastatel 2009 – 2016 kasutatud ka Geenivaramu doonarite isiksuse mõõtmiseks. NEO-PI-R küsimustik koosneb 240st küsimusest, millele vastatakse viie palli skaalal. Küsimustiku lühen versioon, NEO-FFI, koosneb 60 küsimusest ning mõõdab suurt viisikut, kuid mitte selle 30 alaskaalat. Mõlemad küsimustikud on omandiõiguslikud, nende omanik on Psychological Assessment Resources, Inc. (Costa and McCrae, 1992).

Varasemate küsimustike üleseitus ja kasutus põhines suuresti agregeeritud koondmõõdikutel, kus suur hulk küsimusi võetakse kokku üheks näidikuks, nõnda nagu NEO-PI-R'i 240 küsimust annavad 30 isiksuste tahku ning NEO-FFI 60 küsimust 5 peamist mõõdet. Uuemal ajal on aga hakatud rohkem tähelepanu pöörama üksikküsimustele endile, kuna on selgunud, et üksikud küsimused ennustavad erinevaid elusündmuseid täpsemini kui koondmõõdikud (Henry and Mottus, 2024).

Sellest paradigmast lähtudes on loodud 100NP küsimustik, mille 198 küsimust mõõdavad nii suurt viisikut kui ka sellest välja jäävaid iseloomuomadusi. Küsimuste valikul on lähtutud nende vastuste stabiilsusest ja elusündmusid ennustavast potentsiaalist ning autorid on näidanud, et valitud üksikküsimused on annavad piisavalt stabiilseid kasutamiseks agregeerimata kujul. (Henry and Mottus, 2024).

Aastatel 2021 2022 täitsid 100NP küsimustiku 77000 geenidoonarit, kell vastustel põhineb ka käesolev uurimus.

3 Ülevaade matemaatilistest ja statistilistest meetoditest

Antud uurimuses on rõhuasetus eelkõige "valge kasti" meetoditel ja mudelitel, mis võimaldavad tulemusi täpselt tõlgendada. Olukordades, kus ainsaks eesmärgiks on näidata seose olemasolu või puudumist valitud tunnuste vahel, on kasutatud ka "musta kasti" mudeleid, mille puhul ennustuste taga olev loogika kergesti tõlgendatav ei ole.

3.1 Lineaarsed mudelid

Peamise mudelina regressioonanalüüside tegemiseks oleme kasutanud lineaarset regressiooni ning selle edasiarendusi. Lineaarsete mudelite üheks suureks eeliseks on nende tõlgendatavus: tulemused saab kirja panna lihtsa matemaatilise valemina, kus uuritav tunnnus avaldub koefitsentidega läbi korrutatud prediktorite summana. Taoline esitamisviis võimaldab peale ennustuste tegemise uute andmepunktide kohta hinnata iga üksiku prediktori mõju uuritavale tunnusele ning teatud piirini tuvastada isegi põhjuslike seoste olemasolu ja struktuuri uuritavate tunnuste seas.

Lineaarne regresiooni matemaatiline esitus avaldub Valemina 1, kus muutujad x_1 kuni x_n on mudeli sisendid ehk sõltumatud muutujad, w_0 on vabaliige, w_1 kuni w_n on sisendite kordajad ning y on mudeli ennustus ehk sõltuv muutuja ja ϵ märgib kõigi mudelis arvesse võtmata muutujate ning ennustamatuse summaarset mõju.

$$y = w_0 + w_1 x_1 + ... + w_n x_n + \epsilon = w_0 + \sum_{i=1}^{n} w_i x_i + \epsilon \tag{1}$$

Lineaarse mudeli treenimine tähendab kordajate w_0 kuni w_n leidmist, mis minimeeriksid mudeli ennustuste erinevust tegelikkusest. Üheks levinuimaks lineaarse mudeli treenimise metoodikaks on tavaline vähimruutude meetod (ingl. *ordinary least squares*), mis minimeerib mudeli summaarset ruutviga (iga ennustuse poolt tehtud vigade ruutude summat), mis on toodud Valemis 2, kus y_1 kuni y_n on tegelikud andmepunktid ning $\hat{y}_1$ kuni $\hat{y}_n$ neile vastavad ennustused.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

3.1.1 Ülesobitamise vältimine

Sellisel moel treenitud mudelid on aga üpris tundlikud treeningandmetes leiduvate ekstreemsete väärtuste (ingl. *outlier*) suhtes: kuna ekstreemse väärtuse korral on mudeli viga suur ning vea ruut seega eriti suur, kipub juba väike hulk ekstreemseid väärtuseid mudelit liiga palju kõrvale kallutama, mille tulemusena tekib ülesobitamine (ingl. *overfitting*): mudel sobitub küll väga täpselt treeningandmetele ja seal leiduvale juhuslikule mürale,

kuid ei suuda üles leida andmeid genereeriva protsessi tegelikku loogikat ning ei üldistu seega hästi uutele andmetele.

Selle probleemi lahendamiseks kasutatakse kahte peamist taktikat regulariseerimist – väiksemate kaalude eelistamist – ning RANSAC'it ehk ebatavaliste andmepunktide mudelist välja jätmist (Raschka and Mirjalili, 2019).

Regulariseerimie puhul lisatakse optimeerimiskriteeriumiks olevale kaofunktsioonile täiendav liige, et soodustada väikeseid parameetrite väärtuseid.

Üheks levinud regulariseerimise takitkaks on L1 ehk LASSO[2] regulariseerimine, mis lisab minimeeritavale summaarsele ruutveale treenitud parameetrite absoluutväärtuste summa vastavalt Valemile 3, kus w_1 kuni w_k on treenitud parameetrid ning λ on regulariseerimise tugevust määrav konstant. Regulariseerimisel ei kaasata vabaliiget w_0.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{k} |w_i| \tag{3}$$

Teiseks levinud takitkaks on L2 (ingl. *ridge regression*) regulariseerimine, mis lisab minimeeritavale summaarsele ruutveale treenitud parameetrite ruutude summa vastavalt Valemile 4.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{k} w_i^2 \tag{4}$$

L1 regulariseerimine soodustab osade kaaluda seadmist nulliks, mis eemaldab antud liikmed võrrandist. See omadus on kasulik kui eesmärgiks on võrrandi lihtsustamine ebaolulistest muutujatest vabanemise kaudu. L2 regulariseerimine soodustab küll väiksemaid kaale, kuid ei sea kaale võrdseks nulliga: ka väikese mõjuga muutujad jäävad võrrandisse alles.

L1 ja L2 metoodika üldistuseks on nõtkvõrgu mudel (ingl. *elastic net*), mis sisaldab endas nii L1 kui L2 komponenti, mille tugevust saab mõlemat eraldi määrata hüperparameetrite λ_1 ja λ_2 kaudu nagu näidatud Valemis 5.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{i=1}^{k} |w_i| + \lambda_2 \sum_{i=1}^{k} w_i^2 \tag{5}$$

Ebatavaliste andmepunktide mõju vähendamiseks on regulariseerimise – väikeste kaalude eelistamise – kõrval ka teine lähenemine, RANSAC[3], mis proovib andmepunkte jagada sisemisteks (ingl. *inliers*) ja välimisteks (ingl. *outliers*) ning treenida lineaarse mudeli vaid sisemisi andmepunkte kasutades. Algoritmi tööpõhimõte on iteratiivne:

[2]LASSO on lühend mõistele *Least Absolute Shrinkage and Selection Operator*

[3]RANSAC on lühend mõistele *RANdom SAmple Consensus*

1. vali n juhuslikku andmepunkti sisemisteks andmepunktideks ning treeni mudel neid andmepunkte kasutades;

2. lisa sisemiste andmepunktide hulka kõik muud treeningandmetes leiduvad andmepunktid, mille kaugus sobitatud joonest on väiksem kui ette antud kaugus d ning treeni mudel uutel andmetel;

3. korda sammu 2 mudeli viga enam ei vähene.

Neid samme korratakse üldjuhul palju kordi kasutades erinevaid esialgseid sisemisi andmepunkte ning valitakse väikseima veaga mudel (Raschka and Mirjalili, 2019).

Nii regulariseerimisel põhinevad meetodid kui RANSAC mõjutavad üksnes mudeli treenimise ehk kaalude leidmise protsessi. Treenitud mudel avaldub aga ikka Valemis 1 toodud kujul.

3.1.2 Mittelineaarsete seoste modelleerimine

Lineaarseid mudeleid saab kasutada ka mittelineaarsete seoste modeellerimiseks. Üheks levinud lähenemiseks on andmete eeltöötluse sammude lisamine teisendamaks sisendid kujule, millel on väljundiga lineaarsem seos; levinud meetoditeks on sisendi ruutu (või kõrgemale astemele) tõstmine või sellest logaritmi võtmine, et kirjeldada loomult polünomiaalseid ning eksponentsiaalseid seoseid. Nagu näidatud Valemis 6, saame peale taoliste teisenduste tegemist tavalise lineaarse mudeli.

$$y = w_0 + w_1 x + w_2 x^2 + \epsilon \tag{6}$$

Teiseks võimaluseks on sisendparameetri väärtuste piirkonna jagamine mitmeks piirkonnaks (ingl. *spline transformation*), iga piirkonna sees on seos lineaarne, kuid eri piirkondades on joone kaldenurk (ehk mudeli kaalud) erinev. Neid kahte tehnikat saab omavahel ka kombineerida: igas kiirkonnas võib seos olla kirjeldatud sirgjoone asemel ka kõrgema astme polünoomiga.

Taolised teisendused tõstavad aga mudeli keerukust, mis suurendab ülesobitamise riski. Sestap piirdutakse polünomiaalsete teisenduse juures enamasti ruutude ja kuupidega ega kasutata kõrgema astme polünoome.

3.2 Otsustuspuu ja juhuslik mets

Ülalkirjeldatud sisendparameetrite väärtuste mitmeks piirkonnaks jagamise (ingl. *spline transformation*) paradigmal põhinevad ka otsustuspuu tüüpi mudelid.

Otsustuspuude ning nendest nendel põhinevate mudelite üheks eeliseks on võime saada hakkama ka kategooriliste tunnustega. Enamik mudeleid, sealhulgas lineaarsed mudelid, eeldavad numbrilisel kujul antud sisendandmeid, mistõttu tuleb kategoorilised

andmed teisendada eraldiseisvateks tulpadeks, kus iga tulp kätkeb endas 0 või 1 väärtust näidates konkreetse kategooria olemasolu või puudumist (ingl. *one hot encoding*). Otsustuspuud taolist eeltöötlust ei vaja. Samuti suudavad otsustuspuud ära õppida mitme muutuja vahelisest interaktsioonist tulenevad mõjud.

Otsustuspuu toimib sisuliselt nagu linnumääraja: treeningandmetest leitakse kindlad muutujad ja nende kindlad väärtused, mille abil andmepunkte järjest väiksematesse gruppidesse jagama hakata. Puu lehtedes on konkreetsed ennustavad väärtused. Ette antud küsimustele järjest vastates ning seeläbi järjest harusid valides jõutakse lõpuks leheni, milles olev väärtus ongi mudeli ennustus.

Otsustuspuu treenimisel valitakse välja jagamiseks kasutatav tunnus ja selle konkreetne jagamisel kasutatav väärtus nii, et iga tekkiva alamhulga sees oleks ennustatava suuruse väärtus sinna kuuluvatel andmepunktidel võimalikult sarnane.

Otsustuspuu üheks oluliseimaks hüperparameetriks on puu sügavus: liiga väheste otsustussammude korral ei suuda mudel andmeid genereeriva protsessi tegelikku keerukust aduda, liiga paljude otsustussammude korral tekib aga ülesobitumise probleem.

Otsustuspuud on suhteliselt hästi tõlgendatavad, kuna tekkinud puu struktuuri saab graafiliselt välja joonistada ning jagamiskohtadeks valitud tunnused ja nende väärtused on kergesti mõistetavad.

Ennustustäpsuse suurendamiseks ning ülesobitamise probleemide vähendamiseks on levinud suure hulga otsustuspuude treenimine andmestiku erinevatel alamosadel ning nende tehtud ennustuste agregeerimine. Sellist mudelit nimetatakse juhuslikuks metsaks (ingl. *random forest*). Paranenud ennustustäpsuse hinnaks on aga tõlgendatavuse kadu: suur hulk juhuslikul andmestiku fragementidel treenitud puid ei ole enam inimmeelele kergesti mõistetavad, mistõttu ei sobi juhuslik mets olukordades, kus eesmärgiks on andmetes leiduvate seoste mõistmine. Antud töös kasutame juhuslikku metsa selleks, et hinnata seoste tugevust valitud tunnuste vahel olukordades, kus seose täpsem iseloom meile huvi ei paku.

3.3 Determinatsioonikordaja r^2

Üks levinumaid regressioonmudelite headuse ja seoste tugevuse mõõdikuid on seletatavuse määr (ingl. *coefficient of determination*) tähisega r^2, mis näitab, kui suur osa ennustatava muutuja variatiivsusest on võimalik ära seletada ennustava(te) muutuja(te) abil. $r^2 = 1$ näitab, et mudel selgitab ära kogu ennustava muutuja variatiivsuse, $r^2 = 0$ puhul on mudeli ennustused keskmiselt sama head kui konstantselt ennustatava muutuja keskväärtuse ennustamine, $r^2 < 0$ puhul on mudel keskväärtuse ennustamisest ebatäpsem. r^2 arvutamisel jagatakse mudeli summaarne ruutviga SSE (ennustuse erinevus tegelikust andmepunktist) ennustatava muutuja kogu variatiivsusega SST (andmepunkti erisus muutuja keskväärtusest) nagu näidatud Valemis 7 (Raschka and Mirjalili, 2019).

$$r^2 = 1 - \frac{SSE}{SST} = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \mu_y)^2} \tag{7}$$

3.4 Dimensionaalsuse vähendamise meetodid

Oleme käesolevas uurimuses kasutanud kahte dimensionaalsuse vähendamise tehnikat: peakomponent analüüsi ja avastuslikku faktoranalüüsi.

PCA eesmärgiks on selgitada ära võimalikult suur osa andmete variatiivsusest kasutades selleks võimalikult vähe mõõtmeid konstrueerides selle jaoks uued muutujad, mis on mõõdetud muutujate lineaarsed kombinatsioonid. Loodud muutujad on üksteise suhtes ortogonaalsed ega ole seega korreleeritud. Antud töös oleme PCAd kasutanud eelkõige mitmemõõtmeliste andmete visualiseerimiseks kahemõõtmelises ruumis, et leida seaduspärasid ja anomaaliaid. PCA teine kasutusjuht on diskreetsete andmete (näiteks küsitlustulemused skaalal 1-5) muutmine pidevateks tunnusteks, kuna paljud algoritmid töötavad pidevate tunnuste peal paremini kui diskreetsete tunnuste peal.

Avastusliku faktoranalüüsi (ingl. *exploratory factor analysis, EFA*) eesmärgiks on leida üles mõõdetud andmete all olevad varjatud (ingl. *latent*) muutujad. EFA lähtub eeldusest, et mõõdetud muutujate taga peituvad varjatud konstruktid, mida ei olnud võimalik otseselt mõõta, kuid mis on siiski kirjeldatavad ning olulised. Näiteks ei saa otseselt mõõta inimese tervist, kuid sellele saab siiski anda numbrilise hinnangu teiste mõõtude, näiteks kehatemperatuur, pulss, ja põletiku markerid, kaudu. EFA eesmärgiks on nende varjatud muutujate üles leidmine ja võimalikult lihtsalt kirjeldataval kujul esitamine. Saadud muutujad võivad olla omavahel korreleeruvad sõltuvad valitud pööramise (ingl. *rotation*) meetodist. Seetõttu on EFA saadud uued mõõtemd kergemini mõistetavad ja tõlgendatavad kui PCA dimensioonid. PCA on lahendatav EFA teeb andmete kohta vähem eelduseid, eeldab seetõttu iteratiivset lahendamist ning nõuab seetõttu rohkem arvutusvõimsust Osborne and Banjanovic (2016).

EFA esimeseks sammuks on faktorite eraldamine, milleks võib sõltuvalt andmete eripärast kasutada erinevaid meetodeid. Kui andmed alluvad enam-vähem mitmemõõtmelisele normaaljaotusele, peetakse parimaks maksimaalse tõepära (ingl. *maximum likelihood, ML*) meetodit. Kui andmete jaotus erineb oluliselt normaaljaotusest, siis soovitatakse kasutada iteratiivset PAF (ingl. *iterated principal axes factor extraction*) või ULS (ingl. *unweighted least squares extraction*) meetodit Osborne and Banjanovic (2016).

Nii PCA kui EFA puhul on oluliseks sammuks komponentide/faktorite arvu valimine. Kui PCAd kasutatakse abivahendina andmete visualiseerimisel, siis seab koponentide arvule piiri kahemõõtmelisel graafikul kujutada saavate mõõtmete arv. Kui PCAd kasutada vaid pidevate tunnuste loomise eesmärgiga, siis võib alles jätta kõik komponendid, ehk siis sama palju komponente kui oli muutujaid algsetes andmetes. EFA puhul on eesmärgiks aga andmete sisemise struktuuri mõistmine ning seega tuleb faktorite arv

valida nõnda, et see annaks võimalikult hästi tõlgendatavad tulemused. Faktorite arvu valimiseks on välja pakutud mitmeid meetodeid:

- Teoorial põhinevad meetodid – kui andmete kogumise protsess on disainitud eesmärgiga mõõte kindlaid nähtuseid, näiteks suurt viisikut, siis võib see anda hea lähtepunkti faktorite arvu valikule (Osborne and Banjanovic, 2016).
- Kaiseri kriteeriumi (K1) kohaselt tuleks alles jätta kõik faktorid, mille omaväärtus (ingl. *eigenvalue*) on suurem kui 1. Antud riteerium töötab hästi PCA puhul, EFA jaoks on sellest tehtud modifikatsioon, mida nimetatakse minimaalse omaväärtuse kriteeriumiks, mille puhul lävend võib olla madalam kui 1, kuna EFA arvestab ka andmetes sisalduva müraga ning üritab selgitada vaid muutujate jagatud variatiivsust, mitte kogu variatiivsust nagu PCA (Osborne and Banjanovic, 2016). Kaiseri kriteerium kipub faktorite arvu ülehindama pakkudes faktoite arvuks üldjuhul 1/3 kuni 1/6 muutujate arvust (Ledesma and Valero-Mora, 2007).
- Selgitatud variatiivsuse määr – eesmärgiks seatakse selgitada ära n% andmetes leiduvast variatiivsusest ning selle eesmrägi täitmiseks võetakse minimaalne arv komponente (Osborne and Banjanovic, 2016).
- Küünarnukikurv on heuristiline meetod, kus faktorite omaväärtused kantakse graafikule ning otsitakse graafiku tõusunurga järsu muutuse kohta ning võetakse n esimest faktorit kuni kurvi kohani (Osborne and Banjanovic, 2016). Küünarnukikurvi peamiseks probleemiks on selle subjektiivsus: tihtilugu ei ole graafikult võimalik leida ühte ainsat sobivat kurvi kohta. Samuti kipub see meetod faktorite arvu üle hindama, kuid seda peetakse siiski paremaks meetodiks võrreldes Kaiseri kriteeriumiga (Ledesma and Valero-Mora, 2007).
- Paralleel analüüs põhineb Monte Carlo simulatsioonil, kus faktorite omaväärtuseid võrreldakse genereeritud andmetest leitud väärtustega. Seda meetodit peetakse ülalmainitudtega võrreldes täpsemaks ja robustsemaks, kuid ka arvutuslikult kulukamaks. (Osborne and Banjanovic, 2016; Ledesma and Valero-Mora, 2007).
- Minimaalse keskmise alamosa (ingl. *minimum average partial, MAP*) meetod põhineb muutujate ühise variatiivsuse seletamisel ja jätab alles kõik faktorid, mis on vajalikud ühise variatiivsuse säilitamiseks. Osborne ja Banjanovic (2016) peavad seda potentsiaalselt parimaks meetodiks, kuid rõhutavad, et ka selle tulemusi ei tohiks pimesi usaldada (Osborne and Banjanovic, 2016).

Faktorite parema tõlgendatavuse huvides või EFAle lisada faktorite pööramise (ingl. *rotation*). Pööramisel kasutatavad algoritmid jagunevad kahte suurde klastrisse: ortognaalsed (ingl. *orthogonal rotation*) algoritmid hoiavad faktorid mittekoreleerituna (nagu need peale esmast eraldamist on), kaldmeetodid (ingl. *oblique rotation*) lubavad faktoritel

ka korreleeruda, kusjuures enamasti saab lubatud korrelatsiooni määra parameetritega seadistada. Levinumad pööramisviisid on

- *varimax* on ortogonaalne pööre, mis proovib iga faktori laadumised jagada suurteks ja väikesteks ning pöörata faktorit nii, et suured võimalikult palju kasvaksid ja väikesed võimalikult palju kahaneksid;
- *quartimax* on ortogonaalne pööre, mis proovib laadida iga muutuja ühele ja ainult ühele faktorile;
- *equamax* on ortogonaalne pööre, mis kombineerib endas *varimax*'i ja *quartimax*'i lähenemise katsudes laadida iga muutuja eri faktorile lükates samal ajal faktorite struktuuris kaalud võimalikult erinevaks (suured suuremaks ja väikesed väiksemaks);
- *promax* on üldiselt kõige enam soovitatud kaldpööre (Osborne and Banjanovic, 2016).

Kaasajal soovitatakse üldiselt kasutada kaldpöördeid, kuna on ebatõenäoline, et mõõdetavad konstruktid tegelikult üksteisega üldse ei korreleeruks. Kui see peaks aga nii olema, annavad ortogonaalsed ja kaldpöörded peaaegu identseid tulemusi, mistõttu ei ole kaldpöörete kasutamisest vale ka taolises olukorras (Osborne and Banjanovic, 2016).

Käesolevas uurimuses olen avastuslikku faktor analüüsi kasutanud selleks, et konkreetse ennustatava elusündusega kõige rohkem seotud küsimustest väike arv kergesti mõistetavaid iseloomujooni leida.

3.5 Gaussi segumudelid

Gaussi segumudelid (ingl. *gaussian mixture models, GMM*) on juhendamata õppe valda kuuluv generatiivne pehme klasterdamise (ingl. *soft clustering*) mudel. Käesolevas töös on GMMe kasutatud vigaste andmete tuvastamiseks.

GMM lähtub eeldusest, et andmed pärinevad k'ist erinevast normaaljaotusest, peale mudeli parameetrite leidmist on iga andempunkti kohta võimalik leida tõenäosus tema kuulumiseks igasse normaaljaotusest.

Otsustasime GMMi kasutamise kasuks, kuna võrreldes vahest levinuima klasterdamise tehnika K-keskmisega (ingl. *k-means*) on GMMidel kolm olulist eelist:

- GMM, nagu kõik pehme klasterdamise algoritmid, ei määra andmepunkti ühte ja ainult ühte klastrisse vaid võimaldab pakub tõenäosuslikku väljundit, mis võimaldab hinnata andmepunkti kuulumise tõenäosust mis iganes klastrisse. See on vigaste andmepunktid leidmise kontekstis oluline, kuna võimaldab muuta lävendit millest alates andmepunkt välja jätta sõltuvalt sellest, kui suureks hinnatakse eri

vigade kaalu. Kui vigase andmepunkti sisse jätmist peetakse väga suure kaaluga veaks, siis võib välja jätta ka andmepunktid, mille kuulumise tõenäosust vigaste andmete klastrisse hinnatakse näiteks 5% protsendile tavalise 50% asemel.

- GMM annab paremaid tulemusi olukordades, kus klastrite mõõtmed on erinevad ning ühes klastris on andmed palju rohkem hajunud kui teises. Antud juhul on alust oletada, et vigased andmed on palju suurema hajuvusega kui õiged andmed, mistõttu on see omadus oluline.
- GMM arvestab klastrite suurusega ning eeldab, et andmepunkt kuulub suurema tõenäosusega suuremasse klastrisse.
- GMM lubab klastritel olla ka ovaalse kujuga, kuna ei mõõdeta lihtsalt kaugust klastri keskmest, vaid arvestatakse hajuvusega igas mõõtmes eraldi.

GMMi üheks puuduseks küsitlusandmete kontekstis on eeldus, et andmed on pidevad, mitte diskreetsed (nagu vastused skaalal 1, 2, 3, 4, 5).

GMM koosneb k'st n-mõõtmelisest normaaljaotusest, millest igaüks on määratud parameetritega μ (keskväärtuste vektor), Σ (kovariatsiooni maatriks) ja π (klastri suhteline suurus), kusjuures klastri suhteliste suuruste summa peab võrduma ühega vastavalt Valemile 8.

$$\sum_{k \in K} \mu_k = 1 \tag{8}$$

Kovariatsiooni maatriksid on esitavad kujul

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2 & \dots & \sigma_{2,n} \\ \dots & \dots & \dots & \dots \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_n \end{pmatrix} \tag{9}$$

kus σ_i on i'nda mõõtme variatsioon ja $\sigma_{i,j}$ on i'nda ja j'nda mõõtme kovariatsioon. Tunnuse x variatsioon leitakse Valemist 10, kus n on vektori pikkus ehk andmepunktide arv, x_i on konkreetse andmepunkti väärtus ning μ on vektori kesväärtus.

$$var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10}$$

Tunnuste x ja y kovariatsioon leitakse Valemist 11, kus n on vektorite pikkus ehk andmepunktide arv, x_i ja y_i on konkreetsete andmepunktide väärtused ning μ_x ja μ_y on vektorite kesväärtused.

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \tag{11}$$

GMM mudeli treenimisel leitakse EM algoritmiga (ingl. *expectation maximisation algoritm*) parameetrite $\mu_1 ... \mu_k$, $\Sigma_1 ... \Sigma_k$ ja $\pi_1 ... \pi_k$ väärtused nii, et Valemis 12 toodud tõepära funktsiooni väärtus oleks maksimaalne. N on andmepunktide arv ehk ridade arv maatriksis X, K on normaaljaotuste arv GMM mudelis, π_k on k'nda jaotuse suhteline suurus, μ_k ja Σ_k on antud jaotuse parameetrid ning $N(x_n|\mu_k, \Sigma_k)$ on tõenäosus, et andmepunkt x_n on genereeritud jaotusest $N(\mu_k, \Sigma_k)$.

$$p(X) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \tag{12}$$

4 Ülevaade lähteandmetest ja eeltöötlusest

Uurimuses kasutatavas kehamassiindeksi andmestikus oli kokku 860081 näitu 211365 inimeselt.

Avastusliku andmeanalüüsi käigus torkas meile silma, et mõned inimesed on langetavud kaalu ebaloomulikult suures ulatuses, mis põhjustas kõrvalekaldeid treenitud mudelite parameetrites. Selgus, et sellel on kaks põhjust:

- osa inimesi olid läbinud maovähendusoperatsiooni;
- osade näitude puhul olid pikkus ja kaal läinud vahtusse, ehk siis pikkuse väljale oli sisestatud kaal ning vastupidi.

Meil õnnestus saada andmed maovähendusoperatsioonil käinute kohta ning seeläbi 1615 inimest edasistest uuringutest välja jätta.

Vahetusse läinud pikkuse ja kaalu mõõtude tuvastamiseks kasutasime Gaussi segumudeleid. Võtsime eelduseks, et täisealise inimese pikkus püsib kogu elu vältel suhteliselt muutumatuna ning suured kõrvalekalded keskmisest pikkusest viitavad vigastele andmetele. Me ei teadnud aga, kui suurel määral korrektsed pikkuse näidud elu jooksul muutuda võivad ning otsustasime selle leida klasterdamise tehnikaid kasutades.

Arvutasime välja iga pikkuse mõõdu erinevuse antud isiku pikkuse keskväärtusest. Tsentraalse piirteoreemi kohaselt peaksid need erinevused keskväärtusest jaotuma normaaljaotuse kohaselt.

Andmete tegelik jaotus on esitatud Joonisel 1. On näha, et andmed koosnevad mitmest erinevast jaotusest. Üks neist paistab olevat suhteliselt väikese standardhälbega normaaljaotus, mis paistab kirjeldavat korrektseid andmepunkte. Teine jaotus, mis kirjeldab vigaseid andmepunkte, paistab olevat lähendatav suure standardhälbega normaaljaotusele.

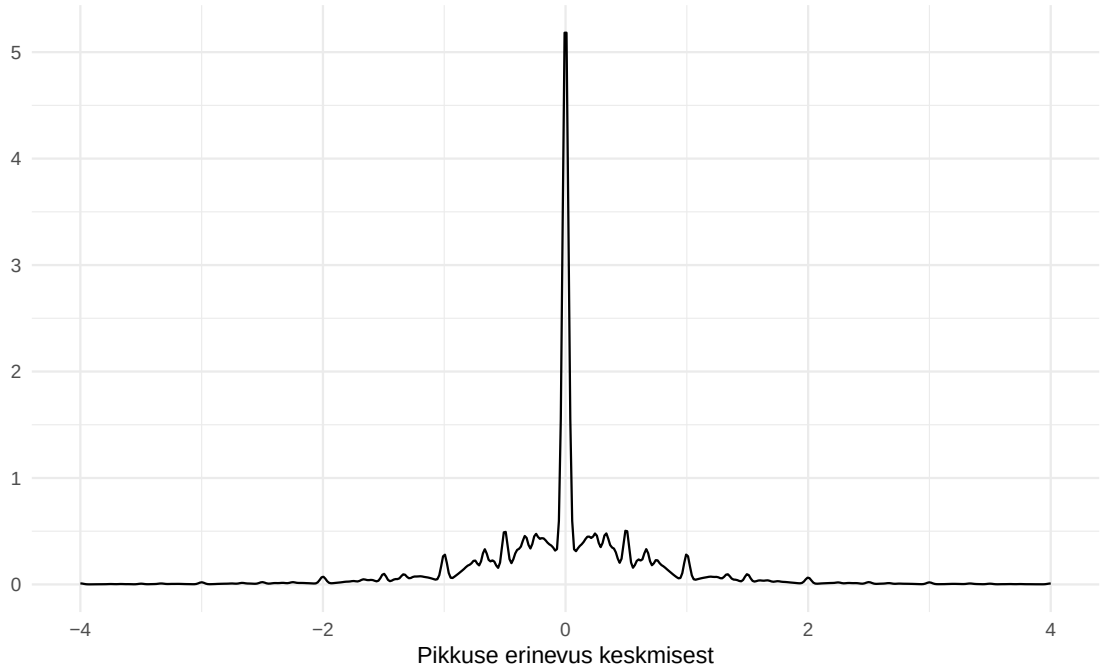
Gaussi segumudelid kinnitasid antud tõlgendust: leidsime, et kõige paremini sobitub andmetele kahe normaaljaotusega segumudel; mõlema jaotuse keskväärtus oli null, väikse standardhälbega jaotus kirjeldas korrektseid andmeid ning laia suure standardhälbega jaotus vigaseid andmeid. Mudeli kohaselt olid mõõtmised korrektsed, kui erinevus keskmisest ei ületanud 1,1 sentimeetrit.

Peale Gaussi segumudelitega tehtud puhastustööd jäi alles 776693 mõõtmist.

100NP küsitluse korrektselt täidetud vastused saime 77174 Geenivaramu doonari kohta. Kaasasime uuringusse inimesed, kelle kohta oli kehamassiindeksi andmeid enne 100NP küsitluse täitmist, jätsime välja peale küsimustikule vastamist kogutud näidud ning jätkasime 45712 uuritavaga.

Tuvastasime inimeste kaalutrajektooridel järgmised olulised punktid, mida selgitab Joonis 2:

- maksimaalne KMI (joonisel punkt α);



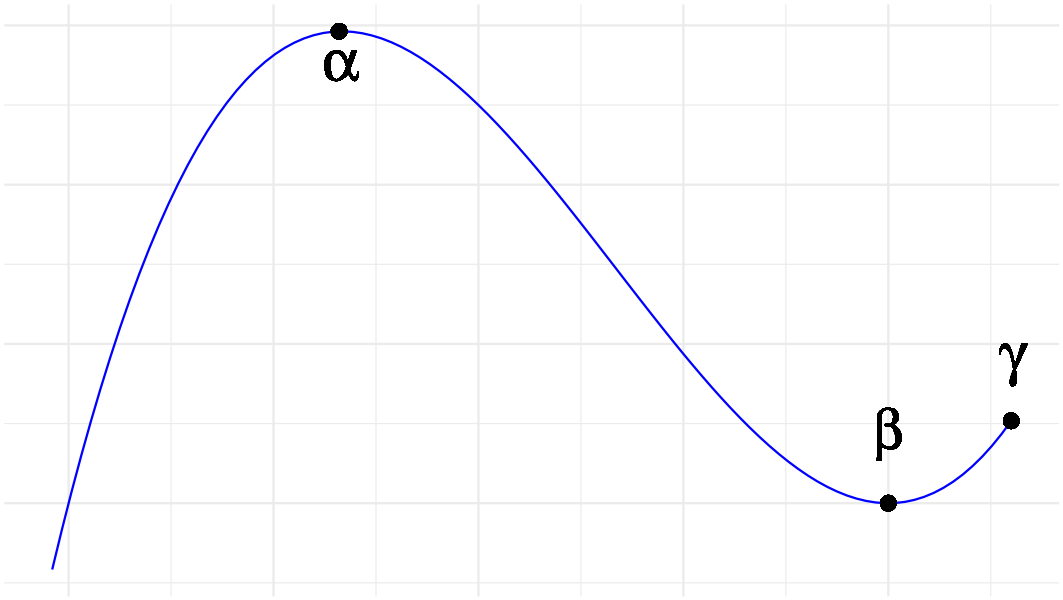
Joonis 1. Pikkuse mõõtude erinevused isiku pikkuse keskväärtusest

- maksimaalsele KMIle järgnenud minimaalne KMI ehk madalaid KMI, mille juurde inimene oli jõudnud peale maksimaalse KMI saavutamist ja sellele järgnenud kaalulangetust (joonisel punkt β);
- hiljutiseim KMI (joonisel punkt γ) ehk viimane KMI näit enne 100NP küsimustiku täitmist.

Defineerisime uuritavate kohta neli KMIga seotud tunnust:

- maksimaalne KMI – maksimaalne KMI mõõtmine enne 100NP küsitlusele vastamist (punkt α);
- hiljutiseim KMI – viimane KMI mõõtmine enne 100NP küsitlusele vastamist (punkt γ);
- kaalulanguse koefitsent – näitab vahet maksimaalse KMI ja maksimaalsele KMIle järgnenud minimaalse KMI näitude vahel $(\alpha - \beta)/\alpha$;
- kaalu taas tõuse koefitsent – näitab vahet maksimaalsele KMIle järgnenud minimaalse KMI ja hiljutiseima KMI näitude vahel $(\gamma - \beta)/\gamma$.

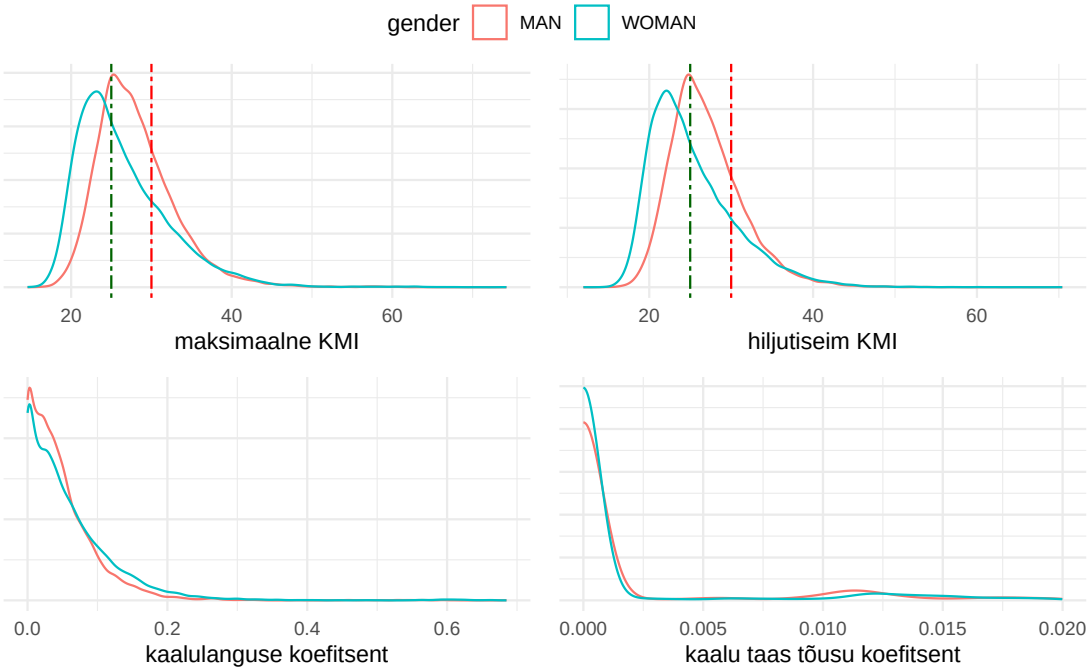
Neid nelja suurust proovime edaspidi iseloomuomaduste põhjal regressioonmudelitega ennustada.



Joonis 2. KMI trajektoori näidis

- maksimaalset KMId ja hiljutiseimat KMId ennustasime kõikide 45712 valimisse kuulunud inimeste puhul;
- kaalulangetuse koefitsenti ennustasime 26170 inimesel, kuna kaasasime vaid need, kelle maksimaalne KMI oli üle 25;
- kaalu taas tõusu koefitsenti ennustasime 10065 inimesel kaasates vaid need, kell maksimaalne KMI oli üle 25 ning kelle kaalulanguse koefitsent oli üle 0.05.

Tunnuste jaotused ülalmainitud valimites on toodud Joonisel 3. Maksimaalne ja hiljutiseim KMI jaotuvad enam-vähem normaaljaotuse kohaselt, kuid neil on positiivses suunas oluliselt pikem saba. Kaalulanguseja kaalu taas tõusu koefitsentide jaotus meenutab eksponentjaotust.



Joonis 3. Kaalu trajektooride tunnuste jaotused. Roheline joon 25 juures märgib ülekaalu piiri, punane joon 30 juures rasvumise piiri. Kaalu taas tõusu jaotusest on välja jäätud ekstreemsed väärtsused.

5 Mineviku KMI ja selle muutuste ennustamine isiksuseomaduste põhjal

5.1 Ennustusmetoodika

Kõik täpsused on leitud viieosalise ristvalideerimisega (ingl. *5-fold cross-validation*), et vähendada treening- ja testandmete juhuslikust jagamisest tekkivat ebastabiilsust.

Kasutasime kolme algoritmi: lineaarset regressiooni, elastset võrku ja juhuslikku metsa. Neist kahe esimese suur eelis on tõlgendatavus: kuna tegemist on lineaarse mudeliega, omistab mudel igale ennustavale muutujale reaalarvulise kaalu, mille märk ja absoluutväärtus võimaldavad hinnata ennustava muutuja mõju suunda ja suurust. Samas teevad need mudelid andmete kohta lihtsustava eelduse oletades, et sisendi ja väljundi vahel kehtib lineaarne seos.

Kuna iseloomuoomaduste puhul ei ole tingimata põhjust seda eeldada, valisime juurde veel juhuslik metsa, mis on tunutud oma täpsuse ja robustsuse poolest ning mida üldiselt peetakse tabeli kujul esitatud andmete puhul üheks parimaks mudeliks. Kuna tegemist on aga keeruka antsambelõppe meetodiga, ei ole see nii kergesti tõlgendatav.

Testisime neid mudeleid erineva suurusega andmestike peal, et uurida, kui palju on andmeid tarvis, enne kui algoritmi täpsus püsiva platoo saavutab.

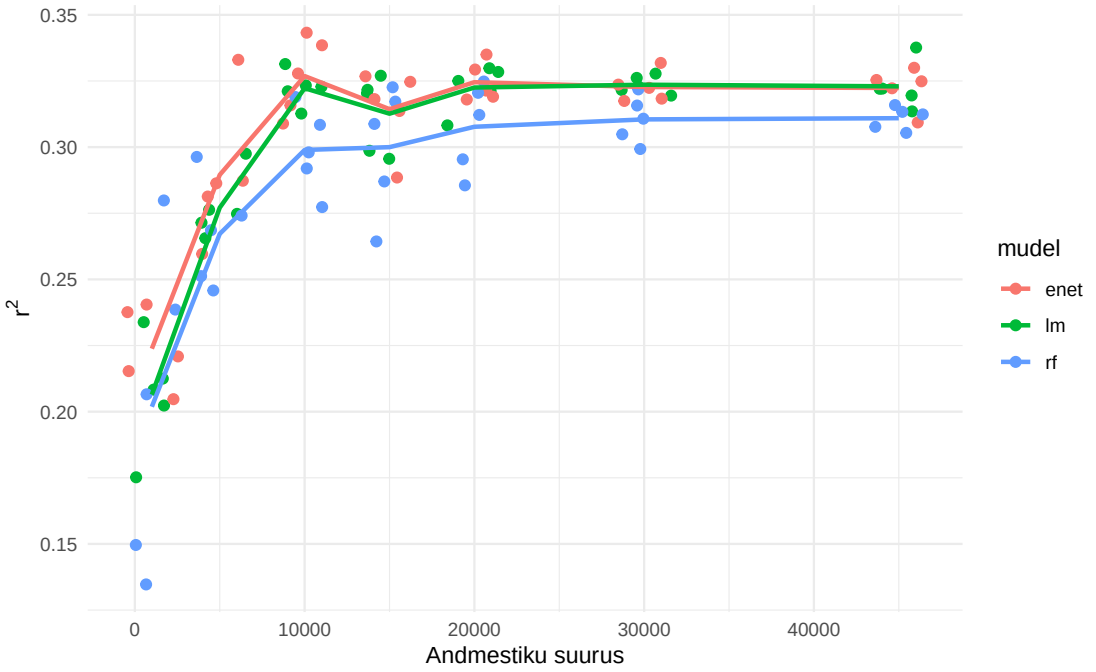
Elastse võrgu mudeli jaoks tuleb ette anda kaks hüperparameetrit: reguleerimise tugevus λ ning L1 ja L2 reguleerimise omavaheline suhe α. Hüperparameetrite väärtused leiti kasutades võre otsingut (ingl. *grid search*) 10000 reaga juhuvalimil.

Soovisime ka kontrollida, kas tulemused erinevad sugude lõikes. Selleks treenisime nii meeste kui naiste jaoks eraldi mudelid võrdse arvu inimeste peal, genereerisime ennustused üle jäänud inimeste jaoks, lahutasime ennustused tegelikest väärtustest ning kontrollisime t-testiga, kas vigade keskväärtused on sugude lõikes erinevad. Ühelgi neljast uurimisküsimusest sugude lõikes erisusi välja ei tulnud.

5.2 Lähimineviku KMI ennustamine

Joonis 4 annab ülevaate antud mudelite r^2 väärtustest eri suurusega valimite puhul. Üksikud punktid näitavad r^2 väärtust viieosalise ristvalideerimise igal iteratsioonil, joon näitab viie iteratsiooni keskväärtust. Elastse võrgu puhul leidisme, et parima tulemuse annavad $\lambda = .2$ ja $\alpha = .1$.

Näeme, et kõikide mudelite tulemused saavutavad platoo juba 20000 inimesega valimi juures. Kui väiksema valimi puhul on regulariseerimist kasutaval elastesel võrgul regulariseerimata lineaarse mudeli ees mõningane eelis, siis 20000st suuremate valimite puhul nende mudelite tulemused võrdsustavad. Huvitav on näha, et üldiselt väga kõrget täpsust näitav juhuslik mets jääb lineaarsetele mudelitele alla kõikide valimi suuruste puhul. Lineaarsed mudelid saavutasid korduvalt $r^2 > .32$, juhusliku metsa parimaks



Joonis 4. r^2 väärtused lähimineviku KMI ennustamisel

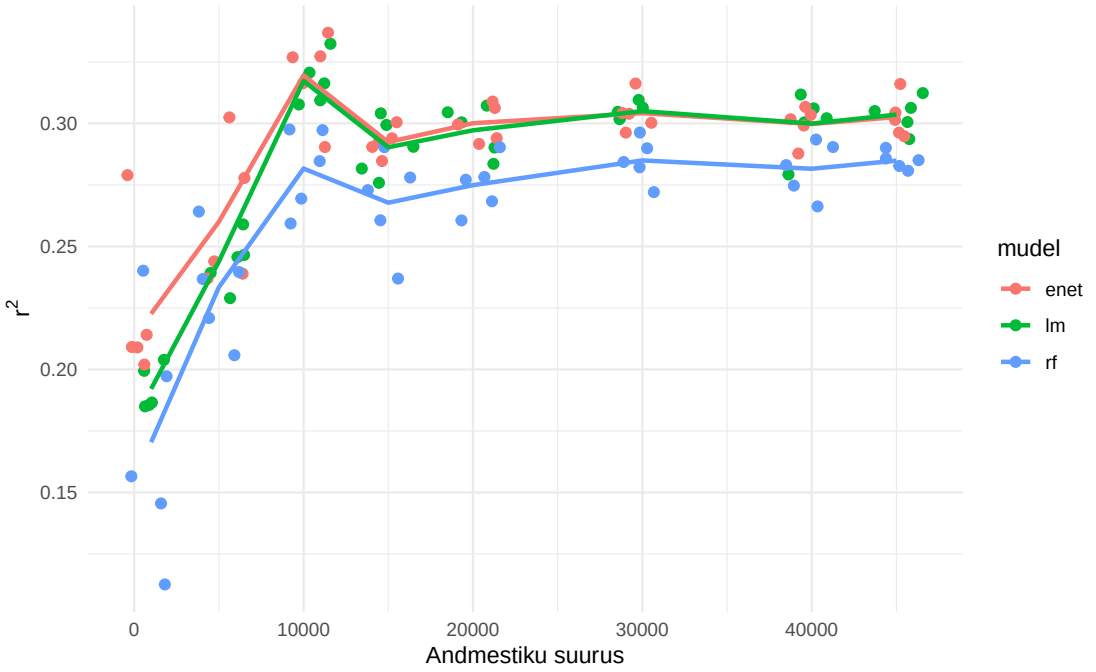
tulemuseks jäi $r^2 = .31$.

5.3 Mineviku maksimaalse KMI ennustamine

Joonis 5 annab ülevaate antud mudelite r^2 väärtustest eri suurusega valimite puhul. Elastse võrgu puhul leidisme, et parima tulemuse annavad $\lambda = .1$ ja $\alpha = .1$.

Üldiselt on tulemused sarnased lähimineviku KMI ennustamisele: mudelite tulemused saavutavad platoo 20000 inimesega valimi juures, regulariseeritud ja regulariseerimata lineaarsete mudelite tulemused on suuremate valimite puhul sisuliselt võrded ning ületavalt selgelt juhuliku metsa tulemusi.

Juhusliku metsa kehvemad tulemused annavad alust oletada, et iseloomujoonte seos KMIga on üpris lineaarne ning et erilist mõju ei oma ka eri iseloomujoonte koosmõjud. Selle hüpoteesi testimiseks treenisime XGBoost mudeli ennustama lineaarse mudeli vigu kasutades samu prediktoreid kui lineaarme mudel ise. Kui iseloomujoonte seos KMIga oleks mittelineaarne, peaks lineaarse mudeli vead olema süstemaatilised ning sestap ennustatavad. Kuna XGBoosti puhul $r^2 \approx 0.01$, samas kui lineaarse mudeli enda puhul $r^2 \approx 0.3$, võib väita, et mittelineaarsete (näiteks U-kujuliste) seoste ning eri iseloomujoonte koosmõju efekt on pigem marginaalne võrreldes lineaarsete efekti proportsiooniga.



Joonis 5. r^2 väärtused maksimaalse mineviku KMI ennustamisel

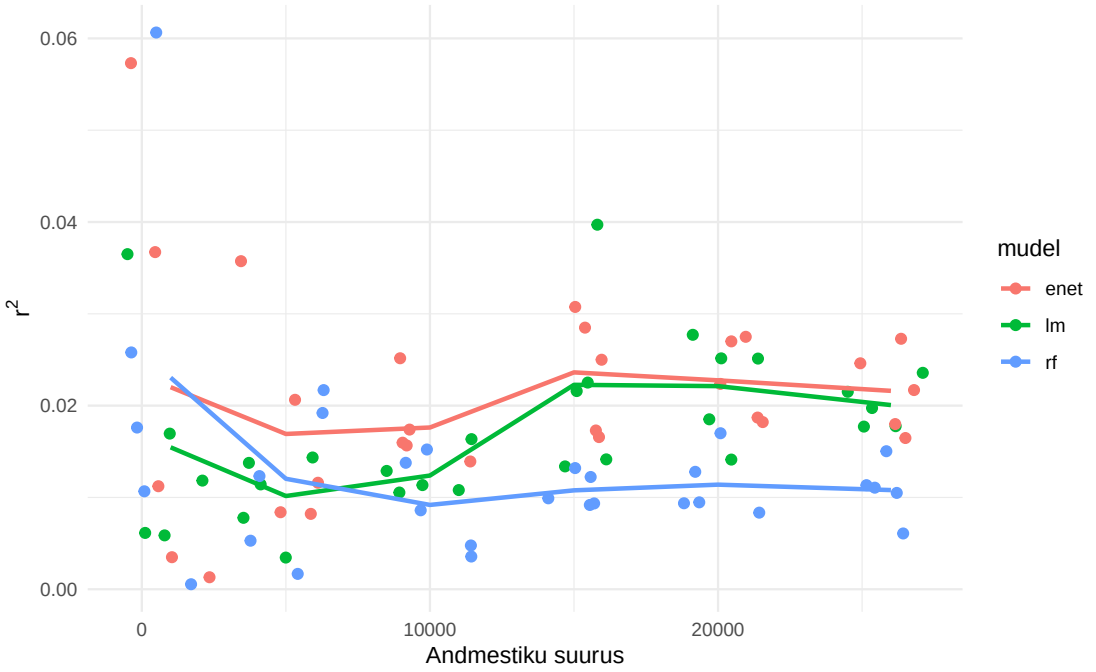
5.4 Mineviku maksimaalse KMI languse ennustamine

Joonis 6 annab ülevaate mudelite r^2 väärtustest eri suurusega valimite puhul. Elastse võrgu puhul leidisme, et parima tulemuse annavad $\lambda = .1$ ja $\alpha = 0$.

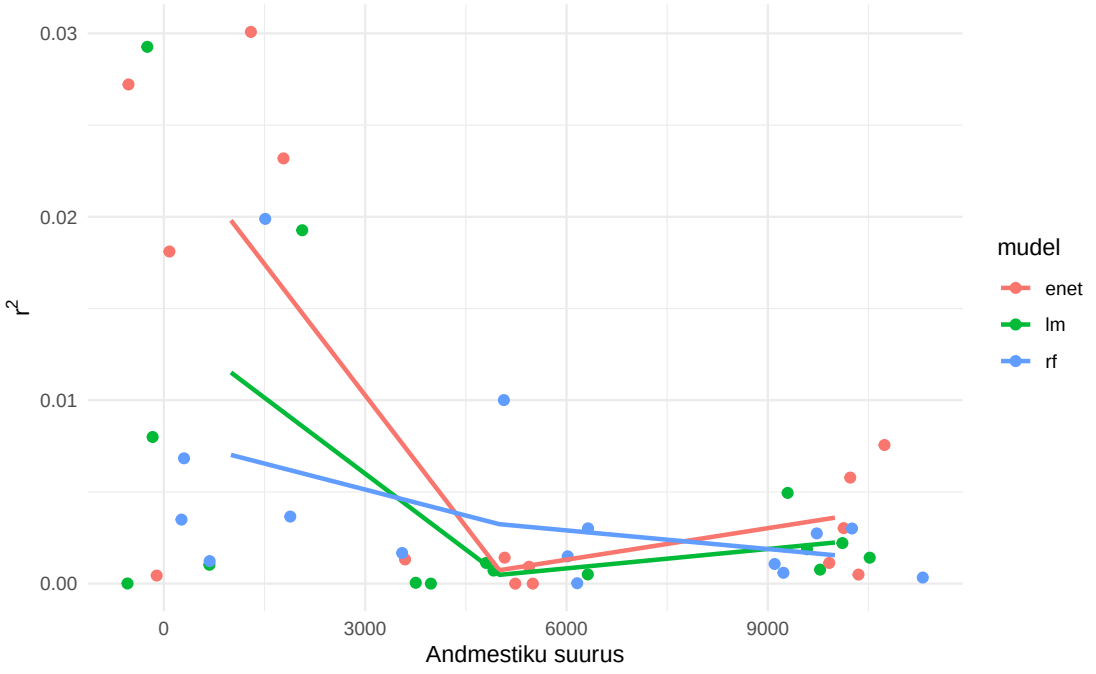
Antud ennustusülesande tulemused on eelmistega võrreldes juba peaaegu suurusjärgu võrra nigelamad, samuti paistab ristvalideerimise iteratsioonide tulemustes olevat rohkem variatiivsust. Kõik mudelid saavutavad platoo juba 15000 inimesega valimi juures ning lineaarsed mudelid on endiselt juhuslikust metsast täpsemad.

5.5 Mineviku maksimaalse KMI taas kasvu ennustamine

Joonis 7 annab ülevaate mudelite r^2 väärtustest eri suurusega valimite puhul. Andmestiku on kaasatud inimesed, kelle maksimaalse KMI on suurem kui 25 ning kes on kaalul lagetanud vähemalt 5% ulatuses võrreldes oma maksimaalse kehakaaluga. Elastse võrgu puhul leidisme, et parima tulemuse annavad $\lambda = .1$ ja $\alpha = 0$.



Joonis 6. r^2 väärtused maksimaalse mineviku KMI languse ennustamisel



Joonis 7. r^2 väärtused maksimaalse mineviku KMI taas kasvu ennustamisel

6 Mineviku KMId ja selle muutusi ennustavad iseloomujooned

Kasutades 100NP iseraporteeritud vastuseid, uurisime, millised iseloomujooned ülalmainitud nelja uurimisküsimusega seostuvad.

Selle jaoks leidsime iga üksikküsimuse ja viie koondfaktori seose tugevuse ja statistilise olulisuse iga uurimisküsimusega treenides ühe muutujaga lineaarse mudeli ja vaadeldes mudeli kordajat. Seejärel valisime välja küsimused, mille mõju oli kõige määravam, ning tegime nendega faktoranalüüsi, et saada väike arv interpreteeritavaid iseloomujooni. Kui statistiliselt olulise seosega iseloomujoonte arv oli suur, siis valisime faktor analüüsi jaoks nendest kõige mõjukama osa küünarnuki kurvi meetodil. Kui statistiliselt olulise mõjuba iseloomujooni oli vähe, kaasasime need kõik faktoranalüüsi. Statistilise olulisuse nivooks võtsime .05 ning mitmese testimise vastu korrigeerisime Benjamini-Hochberg'i meetodil.[4] Faktor analüüsi[5] faktorite arvu määrasime paraleel analüüsi meetodiga[6] ning faktorite keeramiseks kasutasime *oblimin* kaldpööret.

Selgus, et faktorid tulevad sugude lõikes suhteliselt erinevad, sestap tegime kõik protsessi osad alates oluliste küsimuste määratlemisest kuni faktorite pööramiseni mõlema soo jaoks eraldi.

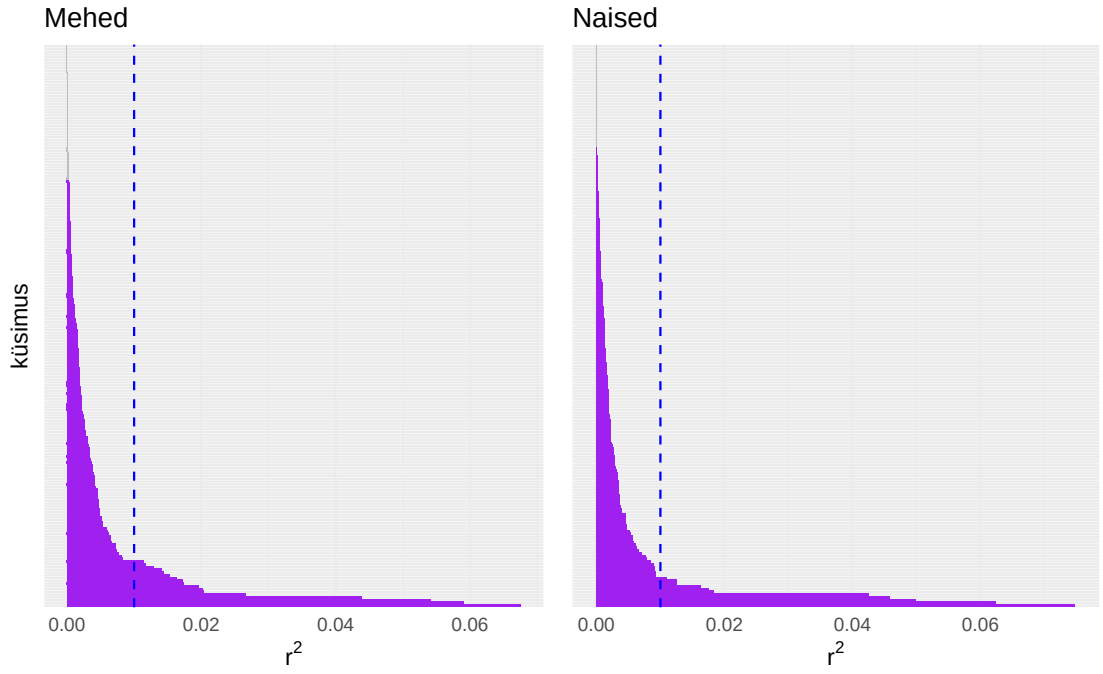
6.1 Maksimaalset KMI väärtust ennustavad küsimused

Küsimuste seosed maksimaalse KMIga on toodud joonisel 8. 204 küsimusest olid meeste puhul statistiliselt olulise seosega 155 ning naiste puhul 167 küsimust. Kuna seoste tugevused varieerusid aga mitme suurusjärgu jagu, otsustasime faktoranalüüsi kaasata vaid küsimused, mille puhul $r^2 > 0.01$, kuna selles kohas oli mõlema soo graafikul kõige selgem kurvi koht. Meestel kaasasime seega anlüüsi 17 küsimust (ära toodud Joonisel 9) ja naistel 11 küsimust (ära toodud Joonisel 10).

Meeste puhul joonistus välja 8 faktorit. Järgnevalt toome ära peamised küsimused iga faktori kohta koos faktoritele antud kirjeldava nimega. Faktori sees on väited esitatud nende kaalu absoluutväärtuse järgi sorteeritult alustades suuremast kaalu absoluutväärtusest. Negatiivse kaaluga väited on tähistatud miinusmärgiga, R'iga lõppevaid väiteid tuleb samuti tõlgendada ümberpööratult.

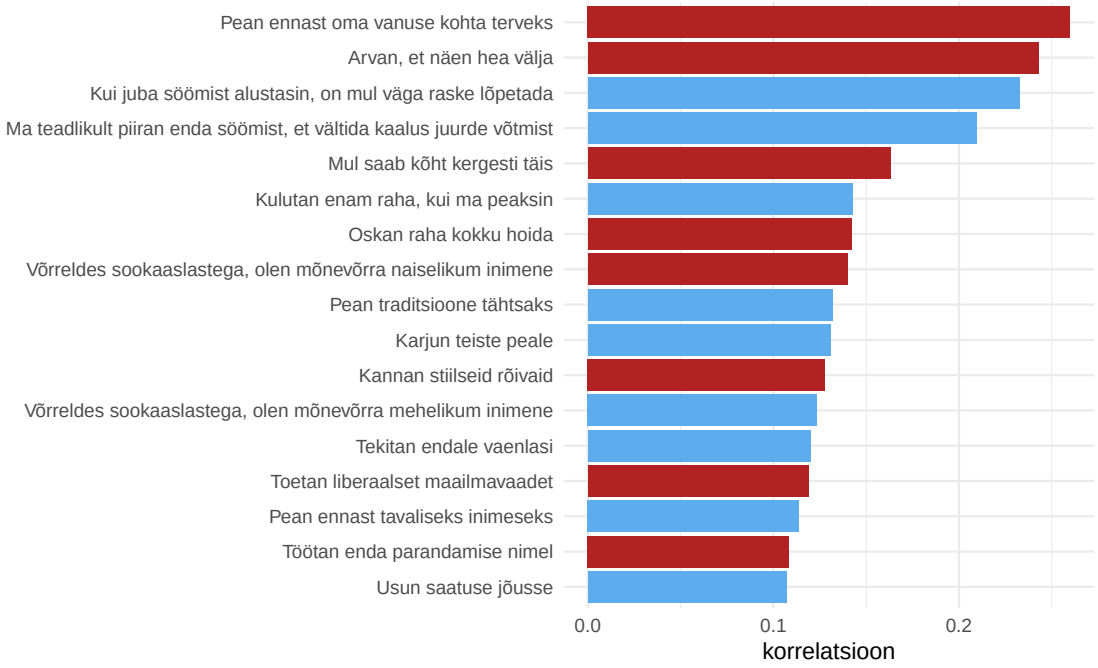
1. raha kulutamine
    - conscientiousness43: Kulutan enam raha, kui ma peaksin
    - -conscientiousness42: Oskan raha kokku hoida

[4] Kasutasime R'i `p.adjust(method="fdr")` meetodit.
[5] Faktor analüüsi jaoks kasutasime R'i paketti `psych`
[6] Paraleel analüüsi jaoks kasutasime R'i paketti `paran`



Joonis 8. Mineviku maksimaalset KMId ennustavad küsimused (lillad küsimused on statistiliselt olulise seosega, hallid küsimused ei ole), sinine punktiir näitab välja valitud küsimuste piiri.

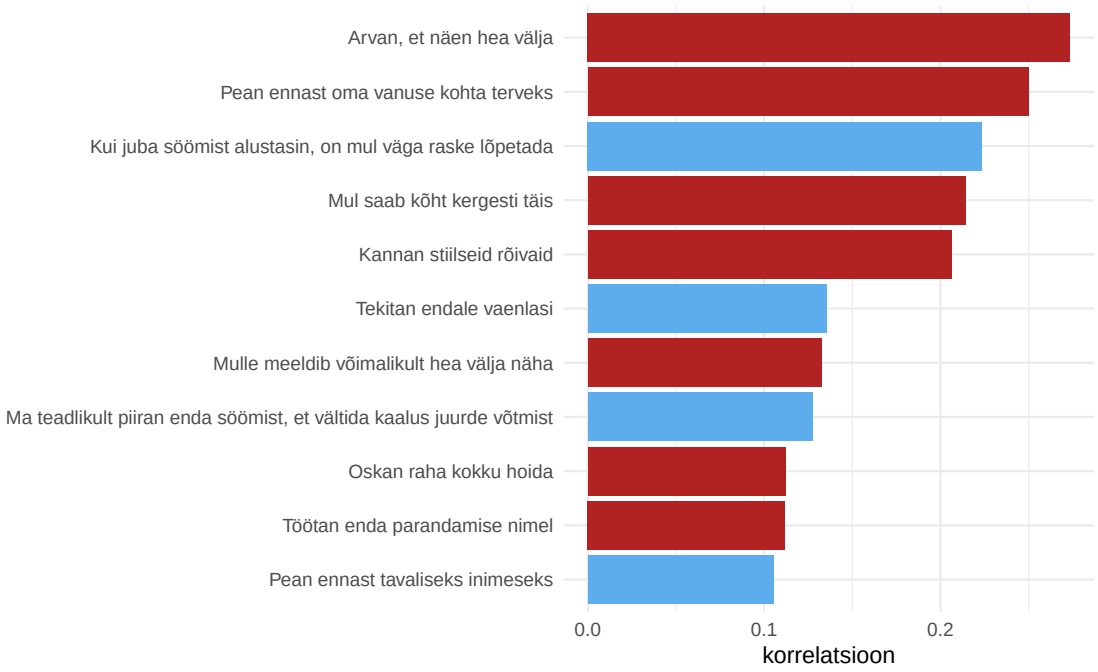
2. söögiisu
    - eat1: Kui juba söömist alustasin, on mul väga raske lõpetada
    - -eat4: Mul saab kõht kergesti täis
3. rahulolu tervise ja välimusega
    - others14: Arvan, et näen hea välja
    - neuroticism43R: Pean ennast oma vanuse kohta terveks
    - others12: Kannan stiilseid rõivaid
4. mehisus
    - others10R: Võrreldes sookaaslastega, olen mõnevõrra naiselikum inimene
    - others29: Võrreldes sookaaslastega, olen mõnevõrra mehelikum inimene
5. sõbralikkus
    - agreeableness19R: Karjun teiste peale



Joonis 9. Mineviku maksimaalset KMId ennustavad küsimused meestel; sinine värv näitab positiivset korelatsiooni (kõrgemat kehakaalu), punane negatiivset korelatsiooni, tulba laius korrelatsiooni absoluutväärtust.

- agreeableness15R: Tekitan endale vaenlasi

6. ambitsioonitus
    - agreeableness36: Pean ennast tavaliseks inimeseks
    - -others12: Kannan stiilseid rõivaid
7. arengu soov
    - -others15: Usun saatuse jõusse
    - openness01R: Pean traditsioone tähtsaks
    - openness03: Toetan liberaalset maailmavaadet
8. distsiplineeritus
    - others27: Töötan enda parandamise nimel
    - eat2: Ma teadlikult piiran enda söömist, et vältida kaalus juurde võtmist
    - other12: Kannan stiilseid rõivaid



Joonis 10. Mineviku maksimaalset KMId ennustavad küsimused naistel; sinine värv näitab positiivset korelatsiooni (kõrgemat kehakaalu), punane negatiivset korelatsiooni, tulba laius korelatsiooni absoluutväärtust.

Naiste puhul joonistus välja viis faktorit:

1. mõõdukus
    - -eat4: Mul saab kõht kergesti täis
    - -conscientiousness42: Oskan raha kokku hoida
    - eat1: Kui juba söömist alustasin, on mul väga raske lõpetada
2. välimuse väärtustamine
    - others13: Mulle meeldib võimalikult hea välja näha
    - others12: Kannan stiilseid rõivaid
    - others14: Arvan, et näen hea välja
3. enese ja teiste väärtustamine
    - -agreeableness36: Pean ennast tavaliseks inimeseks
    - -agreeableness15R: Tekitan endale vaenlasi

4. söömise piiramine
    - Ma teadlikult piiran enda söömist, et vältida kaalus juurde võtmist
5. rahulolu ja arengu soov
    - -neuroticism43R: Pean ennast oma vanuse kohta terveks
    - others14: Arvan, et näen hea välja
    - agreeableness15R: Tekitan endale vaenlasi
    - -conscientiousness42: Oskan raha kokku hoida
    - others27: Töötan enda parandamise nimel

6.2 Kaalu languse koefitsenti ennustavad küsimused

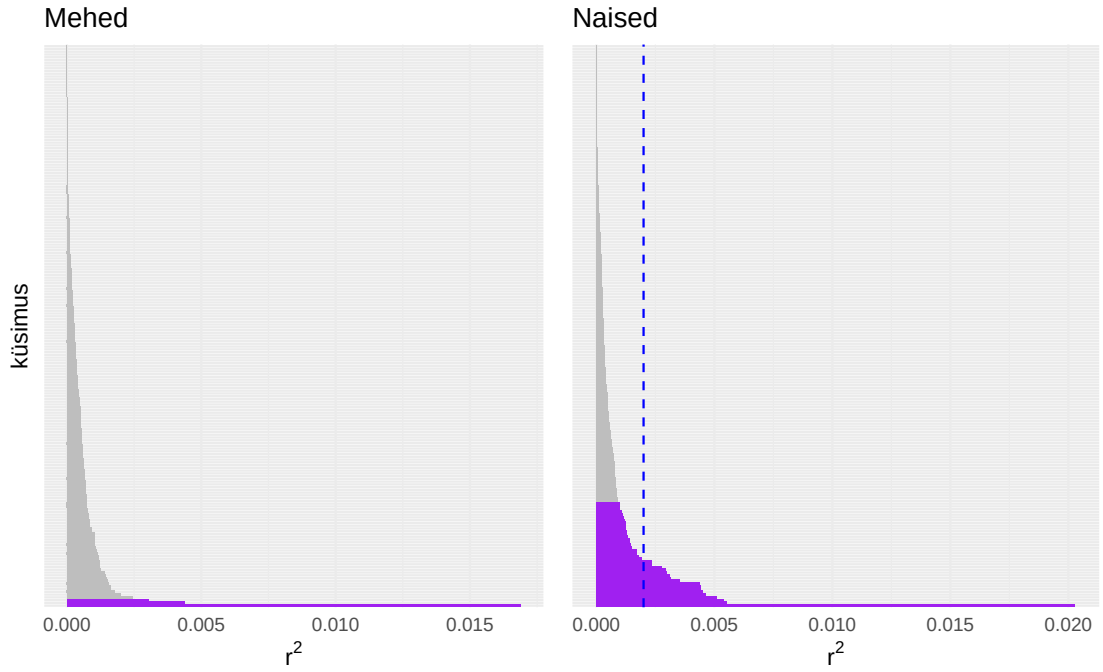
Küsimuste seosed kaalu langusega on toodud joonisel 11. 204 küsimusest olid meeste puhul statistiliselt olulise seosega vaid 3 küsimust (ära toodud Joonisel 12) ning naiste puhul 38 küsimust, mille hulgast valisime 11, mille puhul $r^2 > 0.002$ (ära toodud Joonisel 13).

Mehed:

1. eat2: Ma teadlikult piiran enda söömist, et vältida kaalus juurde võtmist
2. -extraversion23: Mul on liiga palju tegemist
3. others12: Kannan stiilseid rõivaid

Naised:

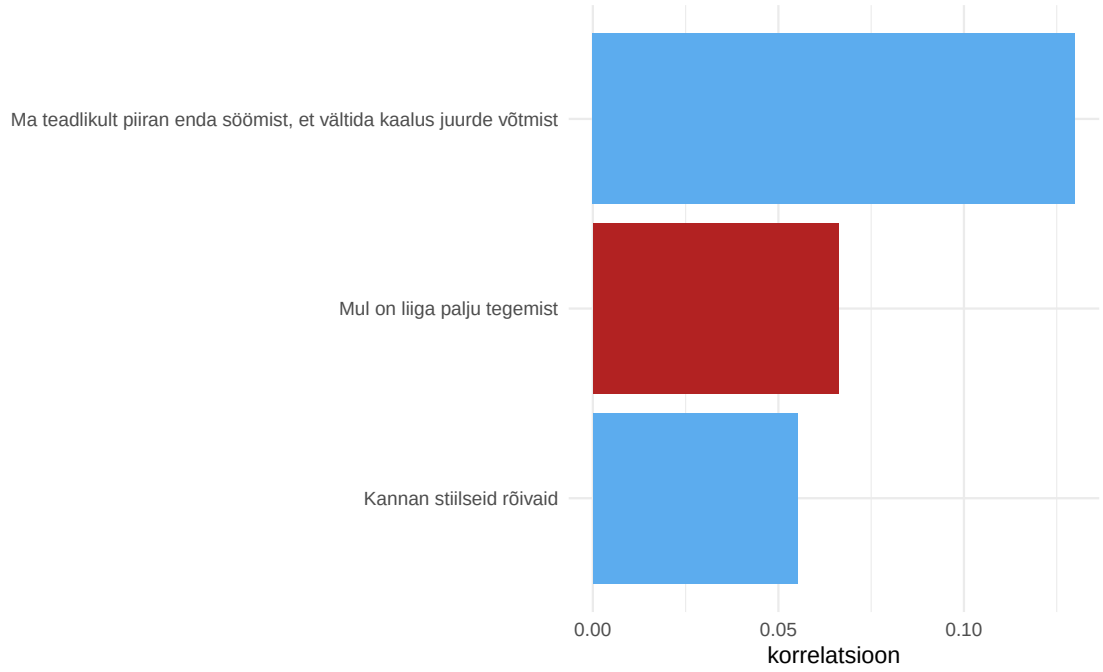
1. sugutung
    - others04: Mul on tugev soov seksi järele
    - others05R: Ma ei mõtle eriti seksist
    - others06: Mulle meeldib flirtida
2. eestvedaja roll
    - extraversion: ekstravertsuse koondskoor
    - extraversion33: Mulle meeldib inimeste hulgas silma paista
    - extraversion12: Tahan olla juhirollis
    - neuroticism43R: Pean ennast oma vanuse kohta terveks
    - -agreeableness49R: Mulle meeldib väga, kui mind tunnustatakse



Joonis 11. Mineviku maksimaalset KMI langust ennustavad küsimused (lillad küsimused on statistiliselt olulise seosega usaldusnivool 0.01, hallid küsimused ei ole), sinine punktiir näitab välja valitud küsimuste piiri

- others06: Mulle meeldib flirtida
- others13: Mulle meeldib võimalikult hea välja näha

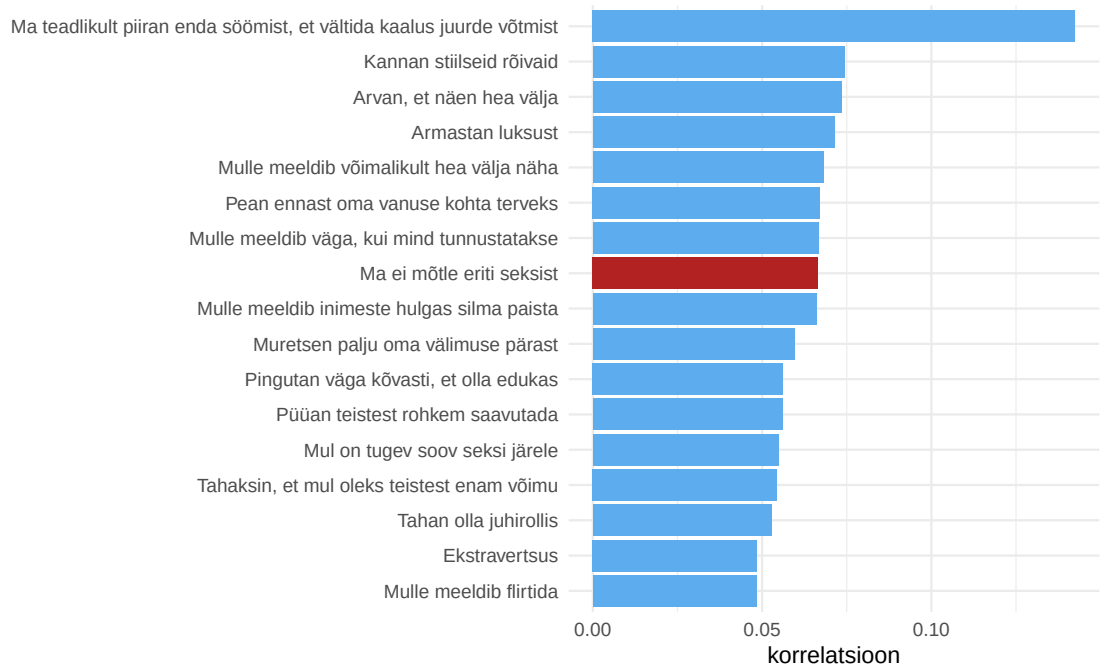
3. ambitsioonikus
    - conscientiousness07: Pingutan väga kõvasti, et olla edukas
    - -agreeableness06R: Püüan teistest rohkem saavutada
    - agreeableness49R: Mulle meeldib väga, kui mind tunnustatakse
4. välimuse väärtustamine
    - others14: Arvan, et näen hea välja
    - others12: Kannan stiilseid rõivaid
    - others13: Mulle meeldib võimalikult hea välja näha
5. ebakindlus välimuse osas
    - neuroticism42: Muretsen palju oma välimuse pärast



Joonis 12. Mineviku maksimaalset KMId ennustavad küsimused naistel; sinine värv näitab positiivset korelatsiooni (kõrgemat kehakaalu), punane negatiivset korelatsiooni, tulba laius korelatsiooni absoluutväärtust.

- -others13: Mulle meeldib võimalikult hea välja näha

6. soov juhtida
    - extraversion11: Tahaksin, et mul oleks teistest enam võimu
    - extraversion12: Tahan olla juhirollis
    - extraversion33: Mulle meeldib inimeste hulgas silma paista
    - -agreeableness06R: Püüan teistest rohkem saavutada
7. edevus
    - agreeableness49R: Mulle meeldib väga, kui mind tunnustatakse
    - others1: Armastan luksust
    - others06: Mulle meeldib flirtida
    - extraversion33: Mulle meeldib inimeste hulgas silma paista



Joonis 13. Mineviku maksimaalset KMId ennustavad küsimused naistel; sinine värv näitab positiivset korelatsiooni (kõrgemat kehakaalu), punane negatiivset korelatsiooni, tulba laius korelatsiooni absoluutväärtust.

6.3 Kaalu taastõusu koefitsenti ennustavad küsimused

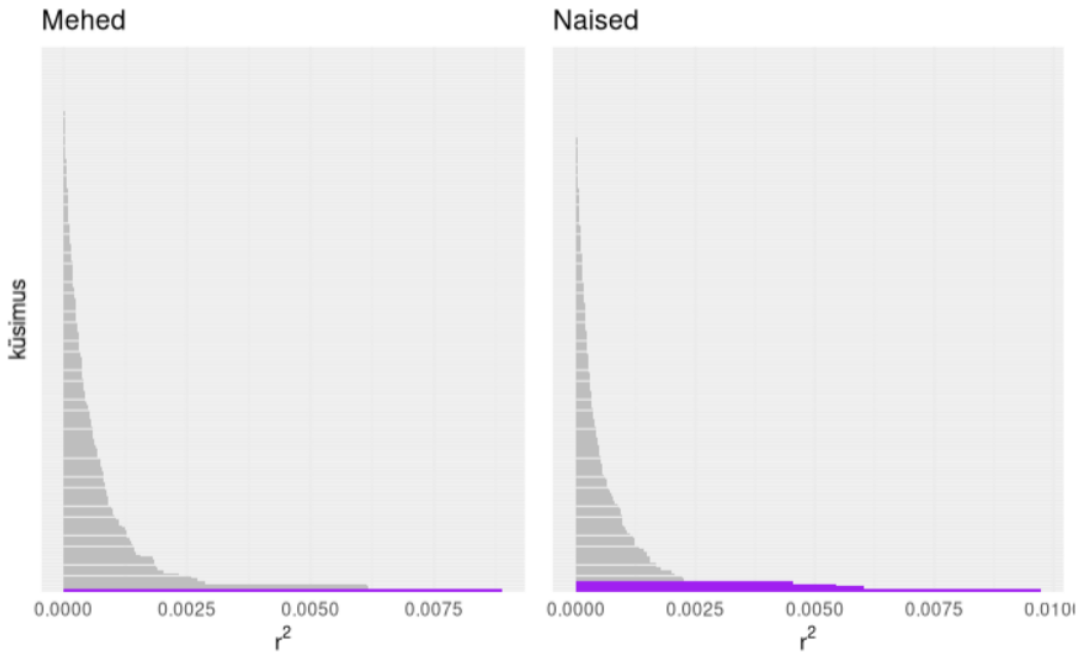
Küsimuste seosed kaalu langusega on toodud joonisel 14. 204 küsimusest olid meeste puhul statistiliselt olulise seosega vaid 1 küsimust: ”Tunnistan kohe, kui olen vea teinud“ (agreeableness40) korrelatsiooniga 0.09 ning naiste puhul 4 küsimust, mis on ära toodud Joonisel 15).

Naiste puhul olulised küsimused laadusid kõik kokku ühte mõõduka söömise faktorisse:

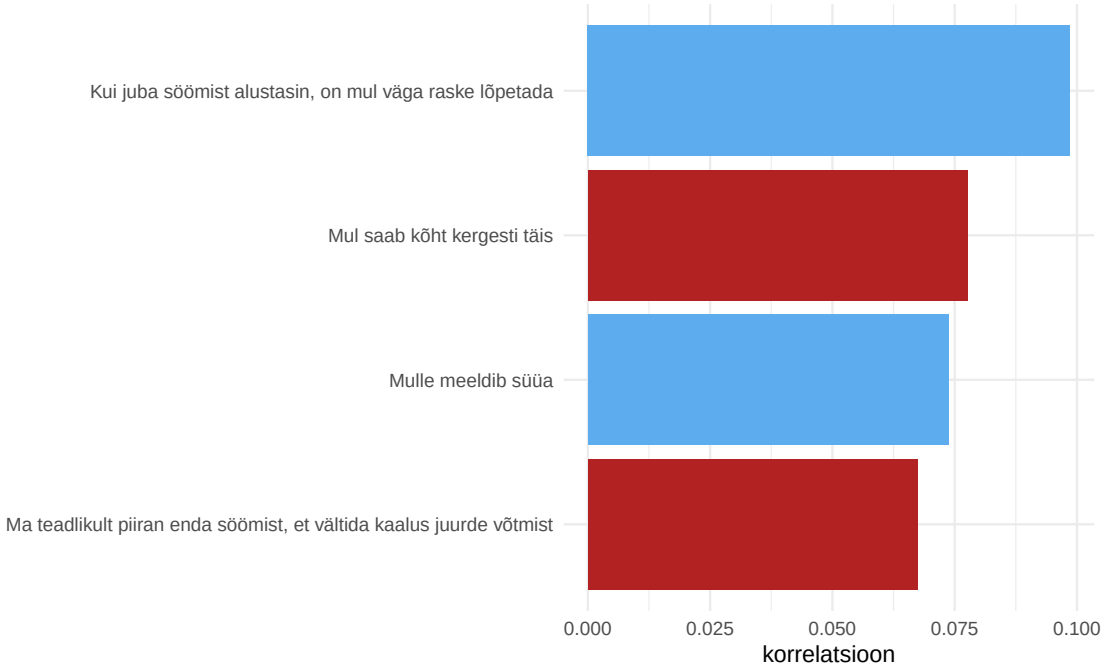
- eat1: Kui juba söömist alustasin, on mul väga raske lõpetada
- eat4: Mul saab kõht kergesti täis
- eat3: Mulle meeldib süüa
- eat2: Ma teadlikult piiran enda söömist, et vältida kaalus juurde võtmist

6.4 Kokkuvõte KMI ja selle muutusi ennustavatest küsimustest

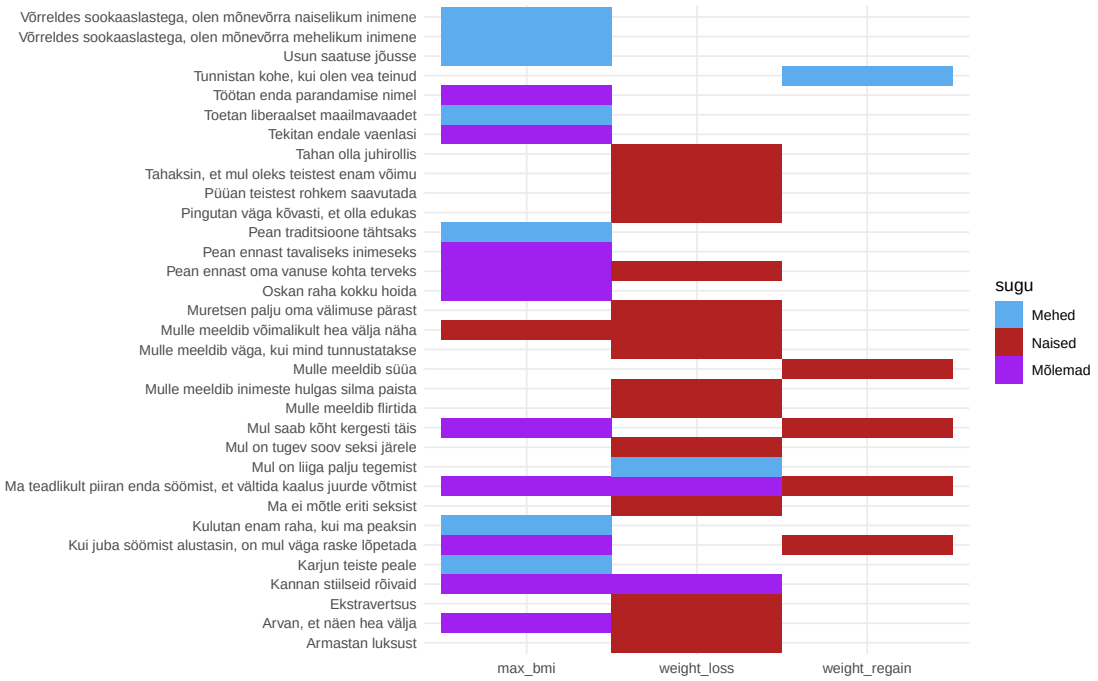
KMId ja selle muutusi ennustavaest küsimustest saab ülevaate Jooniselt 16.



Joonis 14. Mineviku maksimaalset KMI tagasikogumist ennustavad küsimused (lillad küsimused on statistiliselt olulise seosega usaldusnivool 0.01, hallid küsimused ei ole)



Joonis 15. Mineviku maksimaalset KMId ennustavad küsimused naistel; sinine värv näitab positiivset korelatsiooni (kõrgemat kehakaalu), punane negatiivset korelatsiooni, tulba laius korelatsiooni absoluutväärtust.



Joonis 16. Mineviku KMI ja selle muutust ennustavate küsimuste võrdlus. Küsimused, mis on antud tulemi ennustamiseks olulised, on tähistatud värvilise kastiga, kusjuures kasti värv näitab, kas küsimus oli oluline meeste, naiste või mõlema soo jaoks.

7 Arutelu

7.1 KMI ja selle muutuse ennustamine

KMId ja selle muutust ennustavate mudelite võrdluse põhjal võib väita, et isiksuse ja KMI seosed on eelkõige lineaarse loomuga ning sestap hästi kirjeldatavad ja uuritavad lineaarsete mudelitega. Antud väidet tõendab nii juhusliku metsa oluliselt kehvem ennustustäpsus lineaarsete mudelitega võrreldes kui ka asjaolu, et XGBoost ei suutnud lineaarse mudeli vigades leida ette ennustatavat seaduspära.

Selgus ka, et mudelite ennustustäpsus saavutab platoo umbes 20000 treening andmepunkti juures. Sellest edasi ei anna ka regulariseerimise rakendamine eriti suurt täpsuse tõusu.

KMI enda ennustamine iseloomu põhjal annab rohkem kui suursujärgu võrra täpsemaid tulemusi kui KMI muutuse ennustamine: nii maksimaalse kui lähima KMI ennustamisel oli $r^2 \approx 0.3$, KMI languse puhul $r^2 \approx 0.02$ ja KMI taas tõusu puhul $r^2 < 0.005$.

Ühegi ennustusülesande juures ei sõltu ennustuse täpsus statistiliselt olulisel määral soost.

Antud järelduste puhul tuleb meeles pidada, et ennustame iseloomu põhjal minevikus toimunud KMI muutuseid ning samad seaduspärad võivad kuid ei pruugi kehtida tuleviku KMI ennustamise juures.

7.2 KMI ja selle muutusega seotud iseloomujooned

Kaardistades iseloomujoonte seost maksimaalse KMI, KMI languse ja taastõusuga ilmens, kõikidel puhkudel on naistel statistiliselt olulise seosega iseloomujoonte (üksikküsimuste ja suure viisiku koond faktorite) arv suurem kui meestel: maksimaalse KMI puhul olid naiste puhul statistiliselt olulised 167 küsimust, meestel vaid 155. Veelgi enam, KMI languse puhul olid naistel olulised 38 küsimust, samas kui meestel leidus oluline seos vaid 3 küsimuse juures. KMI taastõus seostus naistel nelja ja meestel vaid ühe küsimusega. Samas tuleb arvestada, et valimis oli mehi oluliselt vähem kui naisi, mistõttu on meeste puhul raskem statistilise olulisuse piiri ületada, kui uuritav seos on nõrk.

Näeme, et KMI on seotud pea kõikide inimese iseloomu aspektidega. Antud meetoditega ei ole võimalik mõõta nende seoste suunda, kuid võib oletavad, et vähemalt mingil määral töötavad need seosed mõlemat pidi.

Maksimaalse KMI juures mängis mõlema soo puhul olulist rolli distsiplineeritus, arengule orienteeritud ambitsioonikas mõtteviis ning oma tervise ja välimuse väärtustamine. Meeste puhul ilmnes lisaks veel ka mehisuse faktor, mis suure tõenäosusega tuleneb sellest, et kõrgem KMIi võib peale kõrge rasva protsendi tulla ka suurest lihsamassist, mistõttu on ka pideva jõutreeninguga tegelevatel meestel KMI keskmisest kõrgem.

KMI languse ja taastõusu puhul olid aga välja joonistuvad faktorid sugude lõikes üpriski erinevad.

KMI langus seostub meestel söömise piiramise, välimuse väärtustamise ning rahuliku elutempoga. Naise puhul ilmneb aga lisaks välimuse väärtustamisele hoopis abmitsioonikuse, juhi rollis olemise soovi, seksuaalsete tungide ja tunnustusvajaduse olulisus.

KMI taastõusu puhul olid kõik neli naiste puhul olulist küsimust seotud söömisega, samas kui meeste puhul oli oluline hoopis oma vigade tunnistamise võime.

Taaskord tuleb antud järelduste puhul meeles pidada, et uurime iseloomu põhjal minevikus toimunud KMI muutuseid ning samad seaduspärad võivad kuid ei pruugi kehtida tuleviku KMI ennustamise puhul.

7.3 Saadud tulemuste rakendusvõimalused

Antud uurimuse peamiseks eesmärgiks on aidata kaasa isikupärastatud kaalulangetuse kavade väljatöötamisele. Tänu Geenivaramu andmestikule oleme saanud uurida väga laiahaardelise iseloomujoonte spektri soeseid KMI muutustega suurel ja mitmekesisel inimeste valimil.

Nende seoste kaardistamine aitab mõista, millistele iseloomujoontele tuleb kaalulangetuse edukaks läbimiseks tähelepanu pöörata: ühest küljest aitab see arstil mõista, millised meetodid võiksid antud patsiendi puhul paremaid tulemusi anda ning teisalt aitab leida sobiva metoodika patsiendile, et ta saaks enda kaalulangetust takistavate iseloomuomadustega töötada ning oma vaateid, suhtumisi ja harjumusi muuta.

Patsiendi isiksuseomaduste profiil võiks seega olla arstile üheks personaalse raviplaani koostamise sisenditest ning ideaalis võiks inimene saada isiksuseomaduse küsimustiku täitmise järel teatud soovitusi ka automaatselt.

Selleks, et jälgida inimese iseloomu muutumist ja arenemist kaalulangetuse jooksul, peaks kasutatav küsimustik olema piisavalt lühike, et patsientidel oleks võimalik seda korduvalt täita. Sellised andmed oleksid kasulikud mitte ainult konkreetse patsiendi ravimisel, vaid pakuksid ka hinnalist sisenit edasistele uuringutele isiksuseomaduste muutumise teemal.

Antud töös välja toodud KMIga olulisel määral seotud küsimused ja koondfaktorid võiksidki olla taolise fokuseeritud küsimustiku välja arendamise aluseks.

8 Kokkuvõte

Uurisime iseloomu ja KMI seoseid kasutades rohkem kui 45000 geenidoonari andmeid. Leidsime, et 100NP isiksusetest annab olulist teavet mineviku KMI ennustamiseks ($r^2 > 0.3$), kusjuures KMIga statistiliselt olulises seoses on peaaegu kõik iseloomu nüansid. Iseloomu seosed KMI muutustega on rohkem kui suurusjärgu võrra nõrgemad kui seosed KMI endaga. Iseloomu ja KMI seosed paistavad olevat loomult lineaarsed.

KMI muutustega seostuvad iseloomujooned on sugude lõikes üpris erinevad. Kui mõlema soo puhul tuleb välja distsipliini olulisus ja oma välimuse väärtusamine, siis meeste puhul leidsime seoseid ka elutempo kiiruse ning oma vigade tunnistamise võimega; naiste puhul mängisid rolli aga hoopis ambitsioonikus, juhtimise soov ning seksuaalse ihaga seotud teemad.

Saadud tulemusi saab kasutada lühema fokuseeritud küsimustiku loomiseks, mis aitab arstil panna patsiendi jaoks kokku senisest personaalsem kaalulangetuse kava, et paremini arvestada inimese iseloomust tulenevate iseärasustega. Samuti loob see pinnase meetodite välja töötamiseks, mis aitavad patsiendil muuta oma iseloomu neid aspekte, mis soovitud kehakaalu muutusi takistavad.

Tuleb rõhutada, et kõik käesolaves uurimuses tehtud analüüsid ennustavad iseloomu omaduse pealt mineviku, mitte tuleviku, KMId, kuna tuleviku KMI kohta meil piisavad andemd puudusid. Loodame, et järgmisest uurimustes on võimalik kontrollida siis leitu rakendatavust ka tuleviku KMI ennustamise puhul.

Usume, et paljud siin leitud seaduspärad laienevad peale KMI muutuste ka teistele pingutust nõudvatele positiivsetele elumuutustele.

Viidatud kirjandus

Paul T. Costa and Robert R. McCrae. Revised neo personality inventory (neo pi-r^{TM}) and neo five-factor inventory (neo-ffi). professional manual. 1992.

Michael Gurven, Christopher von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. How universal is the big five? testing the five-factor model of personality variation among foragerfarmers in the bolivian amazon. *Journal of Personality and Social Psychology*, 2013. doi: 10.1037/a0030841.

Sam Henry and Rene Mottus. The 100 nuances of personality: Development of a comprehensive, non-redundant personality item pool. 2024. doi: https://doi.org/10.17605/OSF.IO/TCFGZ.

Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *TODO FILL IN*, pages 114–158. TODO FILL IN, 2008. URL `https://www.ocf.berkeley.edu/~johnlab/pdfs/2008chapter.pdf`.

Laur Kanger. Isiksuseomaduste suur viisik polegi üldkehtiv. 2013. `https://novaator.err.ee/246446/isiksuseomaduste-suur-viisik-polegi-uldkehtiv`.

Sirli Kangur. Meeste ja naiste isiksuseseadumuste erinevused mõõdetuna ncs küsimustiku abil, 2012. URL `https://dspace.ut.ee/server/api/core/bitstreams/73db7480-6512-4b03-bc45-79ff39dac2a3/content`.

Rubén Daniel Ledesma and Pedro Valero-Mora. Determining the number of factors to retain in efa: an easy-to-use computer program for carrying out parallel analysis. 2007. URL `https://www.researchgate.net/publication/241436843_Determining_the_Number_of_Factors_to_Retain_in_EFA_an_easy-to-_use_computer_program_for_carrying_out_Parallel_Analysis`.

Jason W. Osborne and ErinŠ. Banjanovic. *Exploratory Factor Analysis with SAS*. SAS Institute, 2016.

Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning - Third Edition*. Packt Publishing, 2019.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Eerik Sven Puudist**,

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

   Kaalulangetuse edukust ennustavate iseloomujoonte määramine,

   mille juhendajad on PhD Uku Vainik, PhD Raivo Kolde ja PhD Kadri Arumäe.

   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.

Eerik Sven Puudist
15.05.2024