

TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Belinda Lepmets

Pidevate tervisenäitajate kaugeleulatu ennustamine

Bakalaureusetöö (9 EAP)

Juhendaja: Sulev Reisberg, PhD

Tartu 2024

Pidevate tervisenäitajate kaugeleulatuv ennustamine

Lühikokkuvõte:

Selles bakalaureusetöös uuritakse, kas ja kuidas on võimalik luua pidevate tervisenäitajate kaugeleulatuvaks ennustamiseks mudel, mis kasutab lühiajalisi terviseandmeid. Selleks loodi sarnasuse otsimise ennustusmudel, mis põhineb lühiajalise vaatlusaknaga alusandmestikul, milles on paljude patsientide mõõtmiste andmed. Vastavalt lähteparameetritele valitakse kõige sarnasem patsient ning liigutakse samm-sammult mööda tema ja teiste patsientide andmeid edasi, mille tulemusel moodustub kaugeleulatuv prognoos. Seda mudelit demonstreeriti nelja testjuhu peal: pikkus, kehakaal, pulss ja süstoolne vererõhk, kasutades RITA-MAITT ja ELIKTU uuringute andmeid. Parima ennustustäpsusega nendest testjuhtudest on pikkuse ennustusmudel, mille keskmine viga mudeli testimisel oli 5,5 cm ning „kolme ennustuse keskmise“ mudeli puhul 5,1 cm. Mudeli tulemustest võib järeldada, et kuigi kõigi pidevate tervisenäitajate ennustamiseks see ei sobi, annab mõnede näitajate ennustamine piisavalt täpseid tulemusi ning seetõttu väärib meetod edasiarendamist ja uurimist.

Võtmesõnad:

Terviseandmed, ennustamine, pidevad mõõtmised

CERCS:

B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Long-Term Prediction of Continuous Health Measurements

Abstract:

This bachelor's thesis explores the possibility and methodology of creating a model for predicting continuous health measurements over a long-term period using short-term health data. To achieve this, a similarity-based prediction model was developed, relying on a short-term observation window dataset containing measurements from numerous patients. The most similar patient is selected based on initial parameters, and step-by-step progression through their and other patients' data forms the basis for long-term prognosis. This model was demonstrated using four test cases: height, weight, pulse, and systolic blood pressure, utilizing

data from the RITA-MAITT and ELIKTU studies. The height prediction model showed the highest predictive accuracy among these test cases, with an average error of 5.5 cm for model testing and 5.1 cm for the “average of three predictions” model. While this approach may not be suitable for predicting all continuous health measurements, it yields sufficiently accurate results for certain indicators, thus warranting further refinement and investigation of the method.

Keywords:

Health data, prediction, continuous measurements

CERCS:

B110 Bioinformatics, medical informatics, biomathematics, biometrics

P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus.....	5
1. Kirjanduse ülevaade.....	7
2. Metoodika	12
2.1 Mudeli kirjeldus	12
2.2 Testjuhud.....	14
2.3 Lühiajalised andmed mudeli osana	14
2.4 Treening- ja testandmed	15
2.5 Mudeli treenimine ja testimine.....	17
2.6 Tarkvara	18
2.7 Eetikakomitee luba.....	18
3. Tulemused.....	19
3.1 Mudelite parimad kaalud.....	19
3.2 Mudeli täpsuse hindamise tulemused.....	20
3.3 Mudeli ennustuse protsess viie testjuhu näitel	20
4. Arutelu	22
Kokkuvõte.....	26
Tänuõnad	27
Viidatud kirjandus.....	28
LISA 1: Litsents	31

Sissejuhatus

21. sajandi digitaalne ajastu on kaasa toonud revolutsioonilised muutused andmete kogumises ja analüüsis, kaasa arvatud tervisevaldkonnas. Viimaste kümnendite jooksul on elektrooniliste andmete kogumine ja säilitamine muutunud üha tavapärasemaks. See on viinud suure hulga andmeteni, mis kirjeldavad inimeste tervislikku seisundit, harjumusi ja elustiili. Eestis, kus terviseandmete süstemaatilisem kogumine sai alguse 21. sajandi esimese kümnendi lõpus, ei ole see erand. Sellest tulenevalt on nüüdseks kättesaadavad suuremahulised andmekogud, mis võivad ulatuda isegi paari aastakümne tagusesse aega.

Nende elektrooniliste andmete juures on oluline see, et need on kogutud suure hulga inimeste kohta ning ühe ja sama indiviidi kohta võib leida palju eri ajast pärit mõõtmisi. Näiteks võib inimese kehapiikkust ja -kaalu mõõta regulaarselt aastate jooksul. Sellised pidevad mõõtmised annavad võimaluse jälgida muutusi ajas. Lisaks on olemas suur hulk erinevaid mõõdetavaid tunnuseid, mis moodustavad tervikliku pildi inimese tervislikust seisundist.

Kogu selle andmehulga potentsiaal on veel avastamata ja kasutamata, eriti kaugemaleulatuvate ennustuste tegemisel. Näiteks on väljakutseks ennustada inimese kehapiikkust või -kaalu 50 aasta kaugusele, lähtudes vaid lühiajalistest (s.o väiksema kui 20-aastase vaatlusaknaga) andmetest. Üks põhjus selleks on see, et pikemaajaliste ennustuste tegemiseks puudub usaldusväärne meetodika ning andmed ühe inimese kohta võivad olla piiratud lühikesse ajavahemikku.

Bakalaureusetöö eesmärk on uurida, kas ja kuidas on võimalik kasutada lühiajalisi andmeid kaugemaleulatuvate ennustuste tegemiseks, tehes seda inimeste terviseandmete ja pidevate tunnustega nagu kehapiikkus, -kaal, pulss ja vererõhk. Näiteks: milline on isiku pikkus 32-aastaselt, kui ta on praegu 10-aastane ning tema pikkus on 145 cm; milline on isiku kehakaal 60-aastaselt, kui ta on praegu 5-aastane ning tema kaal on 18 kg. Hüpotees on, et selline lähenemine on võimalik ning võib tuua kaasa uusi teadmisi ja praktilisi rakendusi tervisevaldkonnas.

Lisaks uuritakse töös võimalikke piiranguid ja väljakutseid, mis võivad tekkida lühiajaliste andmete kasutamisel kaugemaleulatuvate ennustuste tegemisel. Selgitatakse ka välja, kuidas sellised piirangud võivad mõjutada ennustuste tulemusi ja usaldusväärsust ning milliseid meetmeid saab piirangute leevendamiseks või ületamiseks kasutusele võtta. Samuti tuuakse

töös välja võimalikke lahendusi, mis võivad aidata parandada andmete kvaliteeti ja ennustuste täpsust tervisevaldkonnas.

Töö on jaotatud neljaks suuremaks peatükiks. Esimeses peatükis antakse ülevaade varasematest uurimustest seotud teemadel. Teises peatükis on toodud töös kasutatud meetodite ja töö käigus valminud mudeli kirjeldus. Samuti avatakse töös kasutatud andmestike tausta. Kolmandas peatükis on ülevaade bakalaureusetöö tulemustest ning viimases peatükis analüüsitakse ja arutletakse saadud tulemuste ning edasiarenduste üle.

1. Kirjanduse ülevaade

Juba aastatuhandeid on tuleviku ennustamine inimkonda köitnud [1]. Ennustusmudeleid on tõenäoliselt kasutatud alates sellest, kui inimestel on olnud informatsiooni ja andmeid, aga kaasaegne ennustusmodelite kasutamine võis alata 1940ndatel, kui varajaste arvutite abil hakati analüüsima ilmateadet [2]. Esimesed masinõppe-süsteemid, mis olid kõik keelte tõlkimiseks, loodi 1950ndatel ning Eestiski hakati sellega tegelema sama kümnendi lõpus, kui Tartu Ülikooli matemaatikud tahtsid matemaatilisi tekste tõlkida vene keelest eesti keelde [3]. Tänapäeval, kui on kättesaadavad suuremad kliinilised andmekogud, on tekkinud suurem huvi masinõppe lähenemiste vastu ka tervise valdkonnas [1]. Sellest tulenevalt on populaarseks muutunud kliinilised ennustusmodelid, mida saab kasutada inimese tervisliku seisundi hindamiseks, tervisenäitajate ennustamiseks või ravimeetodite valimisel [4].

Terviseandmetega seotud ennustusmodelite kohta on tehtud mitmeid uuringuid. Näiteks uurisid J. P. Burnham jt [5] 2018. aastal kantavate tehnoloogiate kohta tehtud ennustusmodelite uuringuid. Kantavaks tehnoloogiaks võib lugeda igasugust elektroonilist seadet, mida on võimalik inimese keha peal kanda, näiteks kellad [6]. J. P. Burnham jt [5] leidsid uuringu käigus, et ainulaadseid ning keskmise või hea kvaliteediga ennustusmudeleid, mis olid tehtud inimeste kehal kantavate tehnoloogiate abil ning mis jälgisid tervisega seotud parameetreid ja ennustasid saadud andmete abil suremust või uuesti haigestumist, oli kokku kaheksa. Kuus nendest ennustasid uuesti haigestumist ning kaks suremust. Uuringu järeldusena toodi välja, et kantavate tehnoloogiate abil saadud andmetega tehtud ennustusmodelitel on potentsiaali ennustada kliinilisi tulemusi, kuid see vajab veel hästi kavandatud kliinilisi uuringuid.

Sarnaselt eelnevaga on terviseandmete ennustamismudeleid uurinud D. A. Jenkins jt [7] 2018. aastal ning L. Lapp jt [8] 2023. aastal. Mõlema tööd uurisid dünaamilisi ennustusmudeleid meditsiinivaldkonnast. Mudelit peetakse dünaamiliseks, kui see on disainitud aja jooksul arenema ja lahendama kalibreerimisnihke probleemi ehk saab hakkama rahvastiku demograafia, haiguste levimise, tervishoiusüsteemi jms muutumisega [7]. Mõlema uuringu eesmärk oli teha ülevaade meditsiinivaldkonnas olevatest dünaamilistest ennustusmodelitest, kusjuures L. Lapp keskendus ennustusmodelitele, mis ennustasid patsientide saatust intensiivravi osakonnas. Kuigi huvi dünaamiliste ennustusmodelite väljatöötamise vastu on suur [8], jõudsid mõlemad uuringud järelduseni, et selliseid ennustusmudeleid ei ole siiski väga palju uuritud ja neid on rakendatud ainult piiratud arv tervisenäitajatele (näiteks suremus [8]),

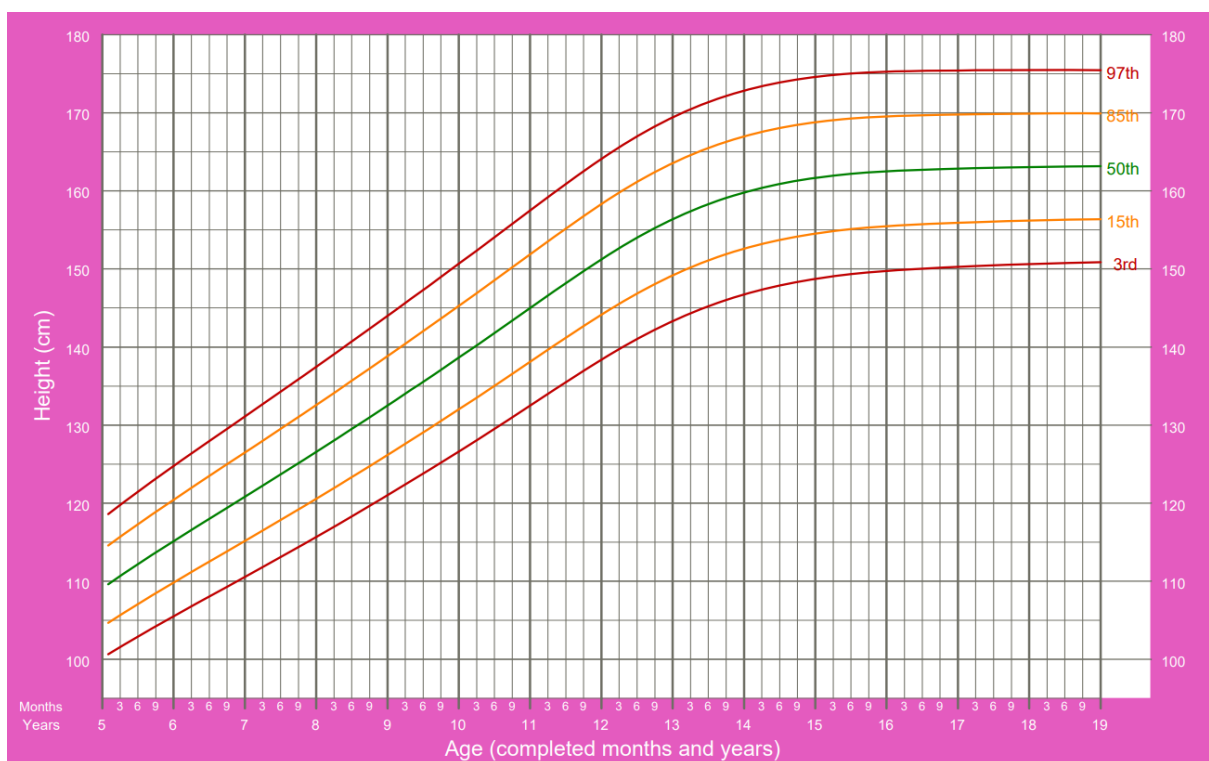
mistõttu oleks kasulik edasi arendada mudeleid, mis keskenduvad ka teistele tõsistele terviseprobleemidele.

Ühe leiutisega tulid 2023. aasta lõpus välja Northeasterni Ülikooli teadlased [9]. Sellest on lähemalt kirjutanud ka Mello-Klein oma artiklis [10]. Ta kirjutab uuest väljatöötatud tehisintellekti tööriistast, mis on kokku pandud transformermudelitest, millele on ette antud väga paljudest elusündmustest saadud teavet, et luua vektorkujutisi *embedding space*-ides, kus mudel hakkab kategoriseerima ja looma seoseid elusündmuste vahel. Mudeli ennustus tehakse samade *embedding space*-ide alusel. Mudel on treenitud kogu Taani rahvastiku terviseandmete peal. See suudab ennustada inimese juures praktiliselt kõike: inimese iseloomust kuni tema elueani, seda täpsusega, mis ületab tipptasemel mudeleid, kasutades selleks elusündmuseid nagu terviseajalugu, haridus, töökoht ja sissetulek. Mello-Klein toob välja, et ühe juhtumuringuna vaadeldi inimeste tõenäosust surra, kus mudeli ennustatud suure surmatõenäosusega inimestest enamik ka surid. Mudeli ennustatud väikese surmatõenäosusega inimeste seas olid surma põhjused enamasti õnnetused, mida mudel ette ennustada ei oska. Samuti on mudel võimeline ennustama indiviidide vastuseid isiksuse küsimustele, eriti kui tegemist on ekstravertse inimesega. Northeasterni Ülikooli teadlased ei ole kindlad, kas seda mudelit saab kasutada ka teiste ühiskondade peal väljaspool Taanit, sest riigi kultuur, seadused ja ühiskonna reeglid mängivad mudeli jaoks suurt rolli.

Selles bakalaureusetöös uuritakse mõõdetavate suuruste – näiteks kehapikkuse – ennustamist. Pikkuse ennustamise meetodite kohta on tehtud palju uuringuid, kuid seda vaid kuni uuritavate 20. eluaastani. Populaarseimad pikkuse ennustamise meetodid on näiteks „kaks aastat korda kaks“, lapse, ema ja isa pikkuste keskmise ehk geenipotentsiaali pikkuse ennustamismeetod ja kasvukõvera meetodid [11,12]. Esimesel neist tuleb lapse täiskasvanuea pikkuse leidmiseks kahega korrutada poisslapse puhul tema pikkus kaheaastaselt ja tüdruklaste puhul teha seda tema pikkusega 18 kuu vanuselt, sest tüdrukud arenevad kiiremini ning seetõttu saab sellega täpsema ennustuse [12]. Seda meetodit kasutakse palju, aga ühtegi uuringut selle meetodi usaldusväärsuse tõestamiseks ei ole tehtud [11,12]. Geenipotentsiaali pikkuse ennustamismeetodit kasutades võetakse ema ja isa pikkuste keskmine ja poisslapse puhul lisatakse keskmisele 6,5 sentimeetrit ning tüdruklaste puhul lahutatakse see [11,13]. Teise võimalusena võetakse isa ja ema pikkuse keskmine, kus poisslapse puhul on eelnevalt ema pikkusele liidetud 13 sentimeetrit ja tüdruklaste puhul isa pikkusest lahutatud 13 sentimeetrit [13,14]. Kasutatakse just 13 sentimeetrit, sest see on keskmine pikkuste vahe mehe ja naise

vahel [13,14]. Arvutuskäigult annavad need kaks võimalust sama tulemuse. Kasvukõvera meetodi puhul tuleb lapse pikkuse ennustamiseks mõõta lapse hetkepikkus ning vaadata kasvukõveralt, millisesse protsentiili laps praeguses vanuses jääb [11]. Täiskasvanuea pikkuse leidmiseks tuleb kasvukõveral mööda sama protsentiili joont või nende vahelt liikuda kuni graafiku lõpuni, saadud tulemus ongi lapse täiskasvanuea ennustatav pikkus [11].

Joonisel 1 on näha Maailma Terviseorganisatsiooni (ingl *World Health Organization*, edaspidi WHO) 2007. aastal välja antud tüdrukute protsentiilideks jagatud pikkuse kasvukõver 5. eluaastast kuni 19. eluaastani. Kui tüdrukulapse pikkus on kuueaastaselt 115 sentimeetrit, siis selle graafiku järgi oleks tema ennustatav täiskasvanuea pikkus umbes 163 sentimeetrit. Lisaks sellele on J. Paley 2004. aastal välja töötanud „korrutustabeli“ meetodi lapse täiskasvanuea pikkuse ennustamiseks, kus pikkuse leidmiseks on vaja lapse hetkepikkus korrutada tabelist võetud vastava kordajaga [15].



Joonis 1. Tüdrukute 5.–19. eluaasta pikkuse kasvukõver protsentiilides [16].

Masinõppe abil on üritanud pikkust ennustada nii J. R. Cordeiro jt [17], M. Shmoishi jt [18], kui ka M. Mlakar jt [19]. Kõigi eesmärk oli luua ennustusmudel, mis ületab täpsusega senised mudelid. Kui kaks esimest nendest üritasid leida olemasolevat regressioonialgoritmi, mis suudab kõige paremini laste täiskasvanuea pikkust etteantud andmete põhjal ennustada, siis Mlakar jt üritasid oma uuringus aga luua täiesti uut algoritmi, mida oma mudelis kasutada.

Cordeiro jt kasutasid oma töös andmestikku, mis sisaldas nii lapse mõlema vanema pikkust, lapse sugu kui ka lapse täiskasvanuea pikkust. Shmoishi jt kasutasid oma töös laste sugu ja nende pikkuste mõõtmisi, mis on võetud kuni kuuenda eluaastani. Cordeiro jt leidsid, et parima tulemuse nende andmetega saavutas masinõppe algoritm LightGBM koos hüperparameetrite häälestamisega (ingl *hyperparameter tuning*), keskmine viga oli 2,1 cm. Shmoishi jt leidsid aga, et nende kasutuses olevate andmetega saadi parim tulemus juhusliku metsa (ingl *Random Forest*) algoritmiga, mille keskmine ennustusviga oli 4–5 cm. Kusjuures mudeli valideerimiseks kasutati kolme erinevat valimi rühma ning kõigi nende täpsus jäi samale tasemele, mis tähendab, et mudelit on võimalik üldistada ka teistele inimrühmadele. Sellest uuringust tuli ka välja, et kõige tähtsamad parameetrid lapse täiskasvanuea pikkuse ennustamisel on lapse sugu ja tema pikkus vanuses 3,4–6,0 eluaastat.

Mlakar jt [19] löid oma uuringus „kasvukõvera võrdlemise“ (*Growth Curve Comparison-GCC*) ennustusmudeli, kus nende andmebaas sisaldas üle 16 000 Sloveenia koolilapse pikkuseid 8. eluaastast kuni 18. eluaastani. Lapse täiskasvanuea pikkuse ennustamiseks kasutab väljatöötatud mudel tema pikkuste aegrida (vähemalt kaks pikkust ja vastavad vanused), aegrea väärtused interpolateeritakse, et need kajastaksid lapse sünnipäeval olevat pikkust. Mudel töötab põhimõttel, et see valib välja sada kõige sarnasema kasvukõveraga õpilast andmebaasist, lapse täiskasvanuea pikkus arvutatakse nende saja õpilase aasta keskmise juurdekasvu summana kuni 18. eluaastani. Uuringu tulemusena saadi, et see mudel on parema täpsusega kui protsentiilide, lineaarse regressiooni, otsustuspuu (ingl *decision tree*) regressiooni ja XGBoost mudelid igas lapse vanuses nii tüdrukutel kui ka poistel. „Kasvukõvera võrdlemise“ meetodi keskmine viga millimeetrites jäi nelja ja 30 vahele, vastavalt lapse viimase pikkuse mõõtmise vanusele. Artiklis oleva graafiku trendijoonelt võib välja lugeda, et mida hilisem on lapse pikkuse mõõtmine, seda täpsem on ennustus.

Kuigi lapse täiskasvanuea pikkuse ennustamiseks on palju meetodeid ja internetist võib leida erinevaid kalkulaatoreid, siis ei ole tehtud meetodit ega kalkulaatorit selleks, et ennustada pikkuseid ja muid tervisenäitajaid 20. eluaastast kaugemale. See on aga vajalik, sest inimesed võivad hakata lühenema juba 30. eluaastast [20,21]. Eluaastatel 30–70 võivad mehed lüheneda ühe tolli (ca 2,5 cm) võrra ja naised isegi kahe tolli (ca 5 cm) võrra [20,22]. Lühenemise põhjuseks võib olla näiteks liigete vaheliste kõhrede kulumine, luude hõrenemine, lihasmassi kadu ning rasvamassi suurenemine [20] või vananemisest tingitud veekaotus lülivaheketastes [21].

Kuigi kõiki vananemisega kaasnevaid protsesse muuta ega aeglustada ei saa ja vähene 2–4 cm lühenemine on normaalne [23], siis liigne lühenemine võib viidata terviseprobleemidele. Seega on oluline tähelepanu pöörata pikkusele ka vanemas eas, mitte ainult lapsena [23], et jälgida, kas lühenemine on ootuspärane või mitte.

Sarnaselt pikkusele võib kirjandusest leida ka kaalu ennustamist inimese elu esimestel aastakümnetel. Samas ei õnnestunud leida ühtki artiklit, mis käsitleks muude tervisemõõtmiste, näiteks pulsi, ennustamist pikemal ajavahemikul.

2. Metoodika

Bakalaureusetöös keskendutakse eestlaste terviseandmete analüüsile ning lühiajaliste andmete põhjal kaugeleulatuva ennustusmodeli väljatöötamisele, millega ennustada mitmesuguseid pidevaid tervisenäitajaid, sealhulgas pikkust, kaalu, pulssi ja vererõhku.

Siin töös peetakse lühiajaliste andmete all silmas samade patsientide kohta eri ajapunktides tehtud sama tervisenäitaja mõõtmiste andmeid. Terviseandmete puhul on need enamasti andmed, mille vaatlusaken on väiksem kui 20 aastat. Selles töös kasutatud andmete korral on see vähem kui üheksa aastat.

Kaugeleulatuvate ennustuste all mõeldakse oluliselt pikemat ajavahemikku, s.o 20–100 aastat.

2.1 Mudeli kirjeldus

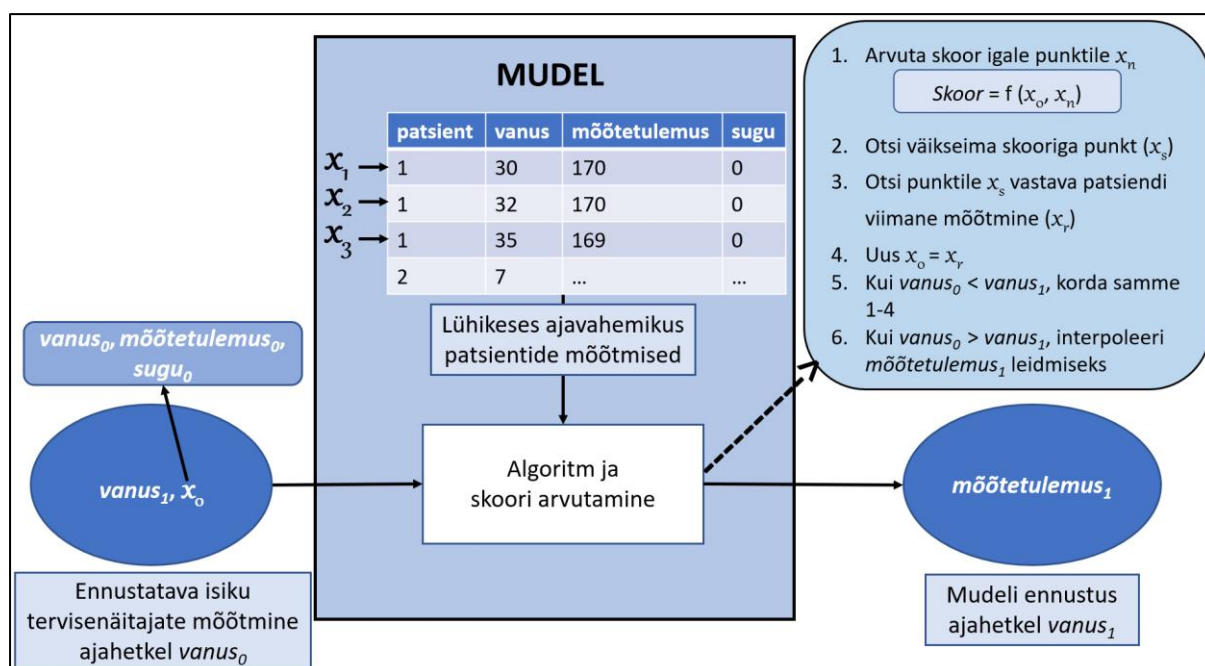
Selles bakalaureusetöös loodud mudel põhineb lühiajalise vaatlusaknaga alusandmestikul, milles leiduvad paljude patsientide mõõtmiste andmed. Vastavalt lähteparameetritele valitakse neile kõige sarnasem patsient ning liigutakse samm-sammult mööda tema ja teiste patsientide andmeid edasi, mille tulemusel moodustub kaugeleulatuva prognoos.

Täpsemalt on tegemist sarnasuse otsimise ennustusmodeliga, mis kasutab algandmestikku, uuritava inimese vanust ($vanus_0$), ennustatavat tervisenäitajat selles vanuses ($mõõtetulemus_0$), sugu ($sugu_0$) ning vanust, milles soovitakse ennustusmodeli abil tervisenäitajat prognoosida ($vanus_1$). Mudel valib $vanus_0$, $mõõtetulemus_0$ ja $sugu_0$ põhjal algandmetest kõige sobivama patsiendi, hinnates nende andmete sarnasust uuritava isiku omadega. Selleks arvutab mudel kõikidele andmestiku mõõtmistele sarnasuse skoori (joonis 2). Skoori arvutusvalem on järgmine:

$$Skoor = a * |\Delta vanus| + b * |\Delta mõõtetulemus| + c * |1/\Delta t| + d * |\Delta sugu|,$$

kus $\Delta vanus$ on patsiendi ja uuritava isiku vanuste vahe, $\Delta mõõtetulemus$ on patsiendi ja uuritava isiku vastava tervisenäitaja mõõteväärtuste vahe, Δt on patsiendi vaadeldava mõõtepunkti ja ajaliselt viimase mõõtepunkti vahe (mida väiksem $1/\Delta t$, seda parem, sest seda pikema ennustuse saab teha), $\Delta sugu$ on patsiendi ja uuritava isiku sugude vahe (0 tähistab sama, ja 1 erinevat sugu) ning a , b , c , ja d on vastavate parameetrite kaalud. $\Delta vanus$, $\Delta mõõtetulemus$ ja

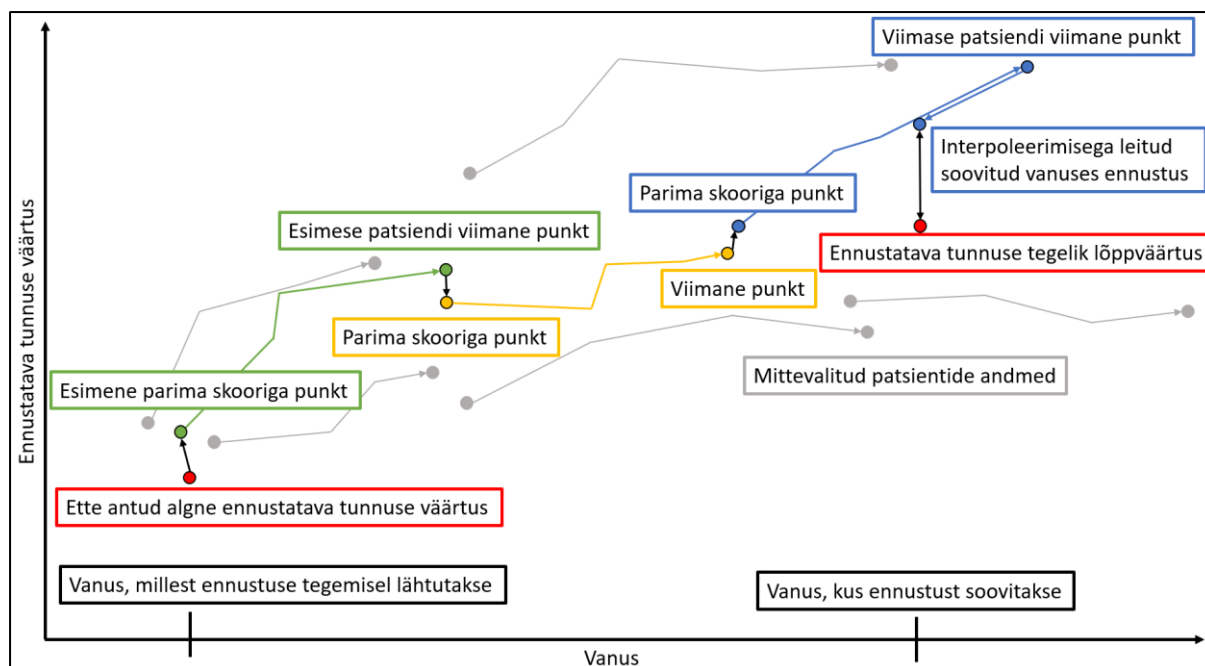
$1/\Delta t$ on normaliseeritud vahemikku 0–1. Skoori kaalud (a, b, c, d) leitakse mudeli treenimise käigus. Mida madalam on mõõtmise skoor, seda sarnasem on see punkt uuritavale isikule.



Joonis 2. Mudeli sisendi ja väljundi kirjeldus.

Kui sobiv (madalaima skooriga) patsient algandmetest on leitud, võtab mudel selle patsiendi ennustatava tervisenäitaja hiliseima mõõtmispunkti ning kasutab selle mõõtmispunkti andmeid, et valida algandmestikust juba omakorda järgmine kõige sarnasem patsient. Mudel jätkab sarnaste inimeste otsimist ja ajas edasi liikumist algandmestiku patsientide peal, kuni jõuab patsiendini, kelle viimase sissekande vanus on võrdne või suurem kui soovitud vanus, mille kohta antud tervisenäitajat sooviti ennustada ($vanus_1$).

Kui patsiendi viimase sissekande vanus on võrdne soovitud vanusega ($vanus_1$), siis on mudel kätte saanud ennustatava tervisenäitaja väärtuse, milleks on selles samas sissekandes ennustatavale tervisenäitajale vastav väärtus, ning lõpetab oma töö. Kui aga patsiendi viimase sissekande vanus ületab soovitud vanust ($vanus_1$), kasutab mudel interpolatsiooni, et leida soovitud vanusele vastav tervisenäitaja väärtus, isegi kui selles vanuses sissekannet patsiendil ei ole. Saadud väärtus ongi ennustustulemus. Kaugeleulatuva ennustuse tegemist kirjeldatud mudeli järgi on visualiseeritud joonisel 3.



Joonis 3. Näide kaugeleulatuva ennustuse tegemisest.

Joonisel 3 antakse mudeli sisendiks ennustatava tunnuse praegune väärtus ja vanus (joonisel punane punkt vasakul). Mudel leiab sellele punktile sarnaseima patsiendi ja seejärel tema viimase mõõtmispunkti (roheline trajektoor). Sellele järgneb sarnaseima patsiendi otsimine (kollane, sinine), kuni saabub vanus, mille kohta sooviti ennustust teha (punane punkt paremal).

2.2 Testjuhud

Selles bakalaureusetöös loodi mudel nelja tervisenäitaja ennustamiseks: kehapikkus (cm), kehakaal (kg), pulss (lööki minutis), süstoolne (ülemine) vererõhk (mmHg). Kõigi tööpõhimõtte on sarnane ja need erinevad üksnes sisendandmetena kasutatud mõõtmiste poolest. Pikkuse ennustamise mudel kasutab lähteandmestikku, kus on olemas patsientide sugu, pikkuste mõõtmised erinevates vanustes ja vanused mõõtmise hetkel ning ka treening- ja testandmetena kasutati üksnes pikkuse andmeid. Kehakaalu, pulsi ja süstoolse vererõhu mudelite puhul oli andmetes kehapikkus asendatud vastava tunnuse mõõtmisega.

2.3 Lühiajalised andmed mudeli osana

Mudeli osana kasutati tervise infosüsteemist pärit terviseandmeid aastatest 2012–2019. Tartu Ülikooli terviseinformaatika uurimiserühm oli andmed puhastanud enne selle lõputöö kirjutamist ja viinud PostgreSQL andmebaasi (RITA MAITT andmestik [24,25]). Sellest

andmebaasist eraldati iga testjuhu jaoks tabeli kujul vajalikud andmed. Tabelid koosnesid neljast veerust: patsiendi ID, patsiendi sugu, patsiendi vanus mõõtmise hetkel ja patsiendi mõõtmise väärtus. Tabelisse jäid ainult nende patsientide kõik mõõtmised, kellel oli algses andmebaasis vastava testjuhu kohta vähemalt neli mõõtetulemust. Nimetatud kriteerium oli rakendatud optimeerimise eesmärgil (võrreldavaid punkte oli tänu sellele vähem). Nelja testjuhu lõplikku mudelis kasutatud tabelite statistikat kajastab tabel 1.

Tabel 1. Statistika lõpliku algandmestiku kohta.

Tervisenäitaja	Kirjete arv	Patsientide arv	Keskmine mõõtmiste arv patsiendi kohta	Min vanus	Max vanus	Keskmine vanus	Min mõõteväärtus	Max mõõteväärtus	Keskmine mõõteväärtus
Pikkus (cm)	59456	10638	5.59	0.01	95.60	38.20	34	207	146.58
Kaal (kg)	73285	12721	5.76	0.00	97.22	38.34	0.6	250	66.35
Pulss (lööki minutis)	24487	4121	5.94	0.12	102.43	59.60	41	184	75.18
Vererõhk (mmHg)	291431	36407	8.00	0.41	107.85	58.93	60	281	135.86

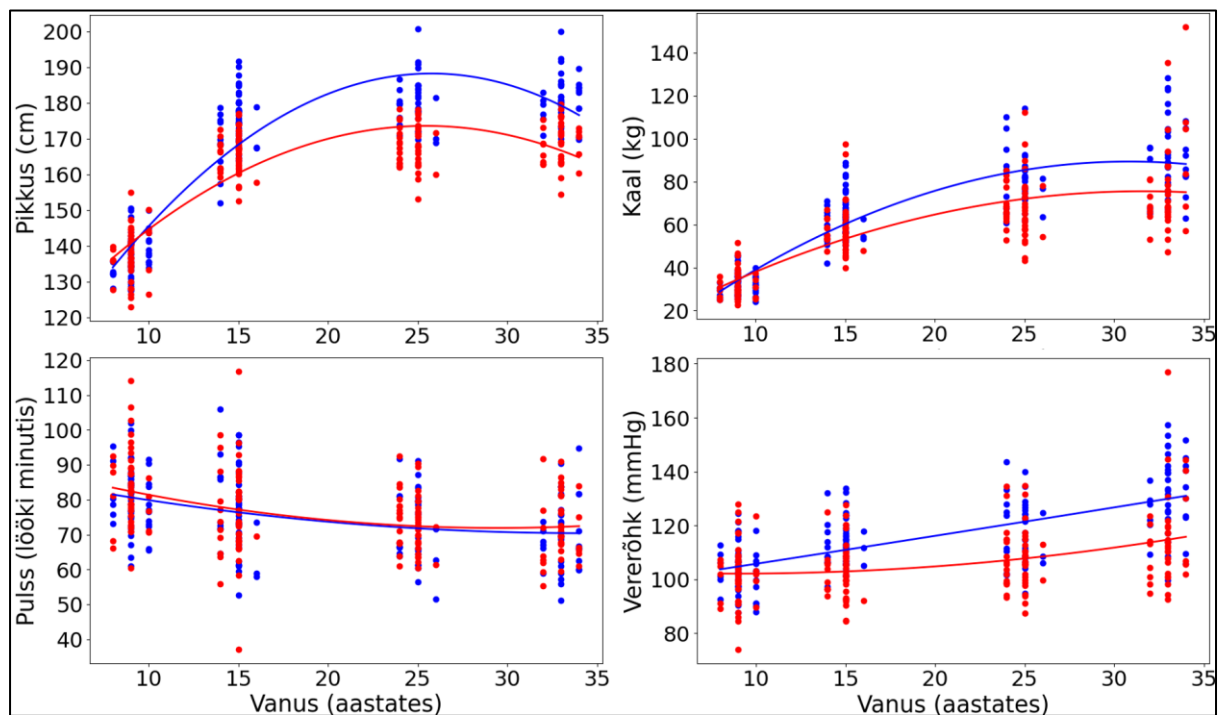
Tabelist 1 on näha, et andmete kogused ja väärtused varieeruvad suuresti kõigi testjuhtude lõikes. Näiteks on pikkuse algandmetes pea 60 000 sissekannet 10 600 patsiendi kohta. Samas kui pulsi algandmestikus on vaid pea 25 000 sissekannet 4 000 patsiendi kohta ja vererõhu algandmetes on üle 290 000 sissekande üle 36 000 patsiendi kohta. Lisaks on näha, et pikkuse ja kehakaalu mudelite algandmestike keskmine vanus on üle 20 eluaasta võrra väiksem kui pulsi ja vererõhu mudelite algandmestikes.

2.4 Treening- ja testandmed

Mudeli treening- ja testandmetena kasutati Eesti laste isiksuse, käitumise ja terviseuuringu (edaspidi ELIKTU) [26] andmeid, millele on lisatud mõned töö autori kogutud andmed.

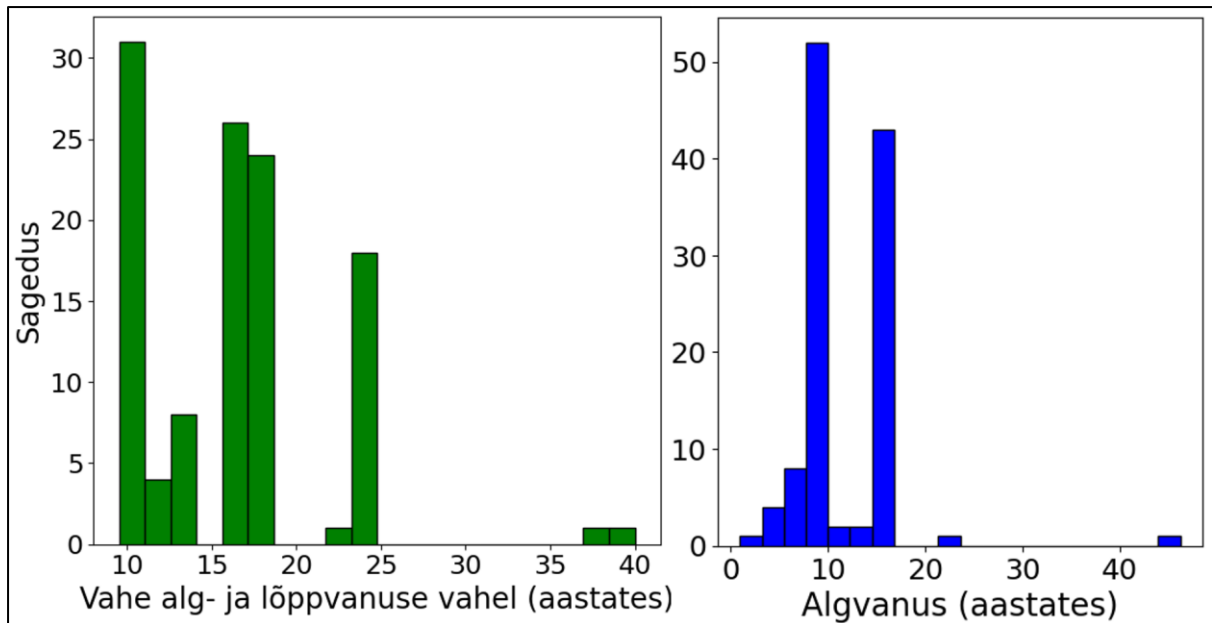
ELIKTU andmestik koosneb sajast inimesest, kelle tervisenäitajaid on kogutud vahemikus 8–34 eluaastat (joonis 4). Nendest sajast inimesest jäi lõpuks töös kasutatud treening- ja testandmestikku 87, sest nendel oli vähemalt kaks sissekannet, mille vaheks oli vähemalt üheksa aastat. Lõplik andmestik koosnes viiest tulbast: sugu, algne vanus, algses vanuses ennustatava tervisenäitaja väärtus, lõppvanus ning tegelik ennustatava tervisenäitaja väärtus

lõppvanuses. Kui inimesel oli algses ELIKTU andmestikus mitu sissekannet, mille vahe oli vähemalt üheksa aastat, siis valiti algne ja lõplik vanus juhuslikult.



Joonis 4. Tervisenäitajate mõõtmised ELIKTU andmetes. Iga punkt tähistab ühte mõõtmist. Pidevad jooned ühendavad iga vanuse keskmisi mõõtmisi (sinine – mehed, punane – naised).

Pikkuse testjuhu andmestikku võeti veel juurde 20 (kümne poisi ja kümne tüdruku) WHO laste kasvukõvera alusel juhuslikult valitud laste algsed ja lõplikud pikkused vanuses 5–19. Lõpuks lisati pikkuse andmestikule veel autori kogutud seitsme inimese andmed, mis teeb pikkuse ennustusmudeli treening- ja testandmestiku suuruseks 114 inimest. Joonisel 5 on näha pikkuse treening- ja testandmestiku andmete jaotust.



Joonis 5. Pikkuse mudeli treening- ja testandmestiku mõõtmiste vanuste ja ajavahe jaotus.

Vasakpoolsel graafikul on näha andmestikus olevate algse ja lõpliku mõõtmise vahet aastates ning parempoolsel graafikul andmestikus olevate isikute algvanuse jaotust.

2.5 Mudeli treenimine ja testimine

Treening- ja testandmestik jaotati juhuslikul valikul 80% treening- ja 20% testandmeteks.

Mudeli treenimise käigus leiti mudelis kasutatava skoori parameetritele parimad kaalud. Selleks kasutati mudelit treeningandmetel 100 erineva kaalude komplektiga, kus iga kord valiti parameetrite kaaluks juhuslik suurus vahemikust 0–1. Iga parameetrite kaalude komplekti läbitegemisel arvutab mudel nii treening- kui testandmete RMSE (ruutkeskmine viga; näitab, milline on tõese ja ennustatud väärtuse keskmine vahe). Lõpliku mudeli kaaludeks valiti parameetrite komplekt, mille RMSE oli treeningandmetel kõige väiksem. Tulemuse illustreerimiseks valiti testandmetest viis juhuslikku inimest ning nende inimeste kohta ennustuse tegemise protsess kujutati joonisel.

Mudeli ennustustäpsuse hindamiseks võrreldi seda ka „andmestiku keskmise“ mudeliga, mis ennustas kõikidele testandmestiku isikutele ühesugust – täpsemalt vastavas vanuses isikute keskmist – mõõtetulemust. Samuti võrreldi seda „kolme ennustuse keskmise“ ennustusmudeliga, mille tööpõhimõte on sama sarnasuse otsimise mudeliga, kuid ainuke

erinevus on see, et esimesel skoorimisel valis mudel kolm kõige sarnasemat patsienti ning lõpptulemuseks saadi nende kolme haru tulemuse keskmine.

2.6 Tarkvara

Selle bakalaureusetöö käigus loodud mudeli tarkvara lähtekood on kättesaadav GitHub repositooriumist lingil <https://github.com/BelindaLep/kaugeleulatuv-ennustamine>.

Repositoorium sisaldab koodi mudeli treenimiseks ja katsetamiseks. Samuti on antud näidis alusandmete ning treening- ja testandmete failid. Täpsemad juhised ja tutvustus on olemas failis „README.md“.

2.7 Eetikakomitee luba

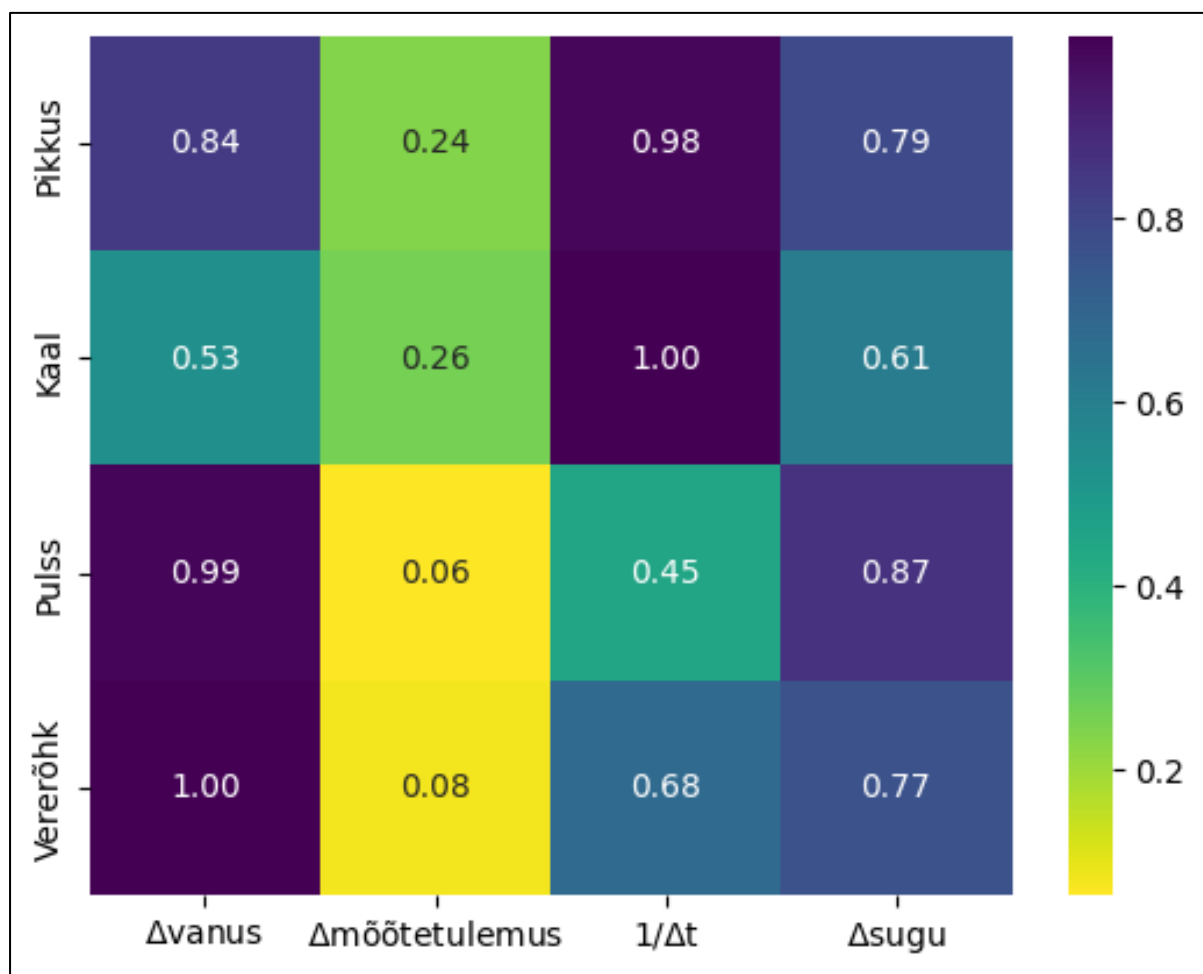
Bakalaureusetöö viidi läbi Eesti bioetika ja inimuuringute nõukogu loa 24.10.2023 1.1-12/3833 alusel ning ELIKTU juhtkomitee loal.

3. Tulemused

Töö raames katsetati mudelit nelja testjuhu peal: pikkus, kehakaal, süstoolne vererõhk ja pulss. Tulemuste osas tehakse ülevaade mudeli testjuhtude parimate kaalude ja täpsuste tulemuste kohta.

3.1 Mudelite parimad kaalud

Joonisel 6 on välja toodud soojuskaart, kus on näha kõigi nelja testjuhu mudeli lõplikud parameetrite kaalud.



Joonis 6. Soojuskaart mudelite kaaludest.

Joonisel on toodud vastav parameeter seda tumedamalt, mida suurema kaalu on see treenimise käigus saanud.

3.2 Mudeli täpsuse hindamise tulemused

Selles töös hinnati mudelite täpsust RMSE abil. RMSE leiti nii mudeli treenimisel kui testimisel. Need tulemused sisestati tabelisse (tabel 2).

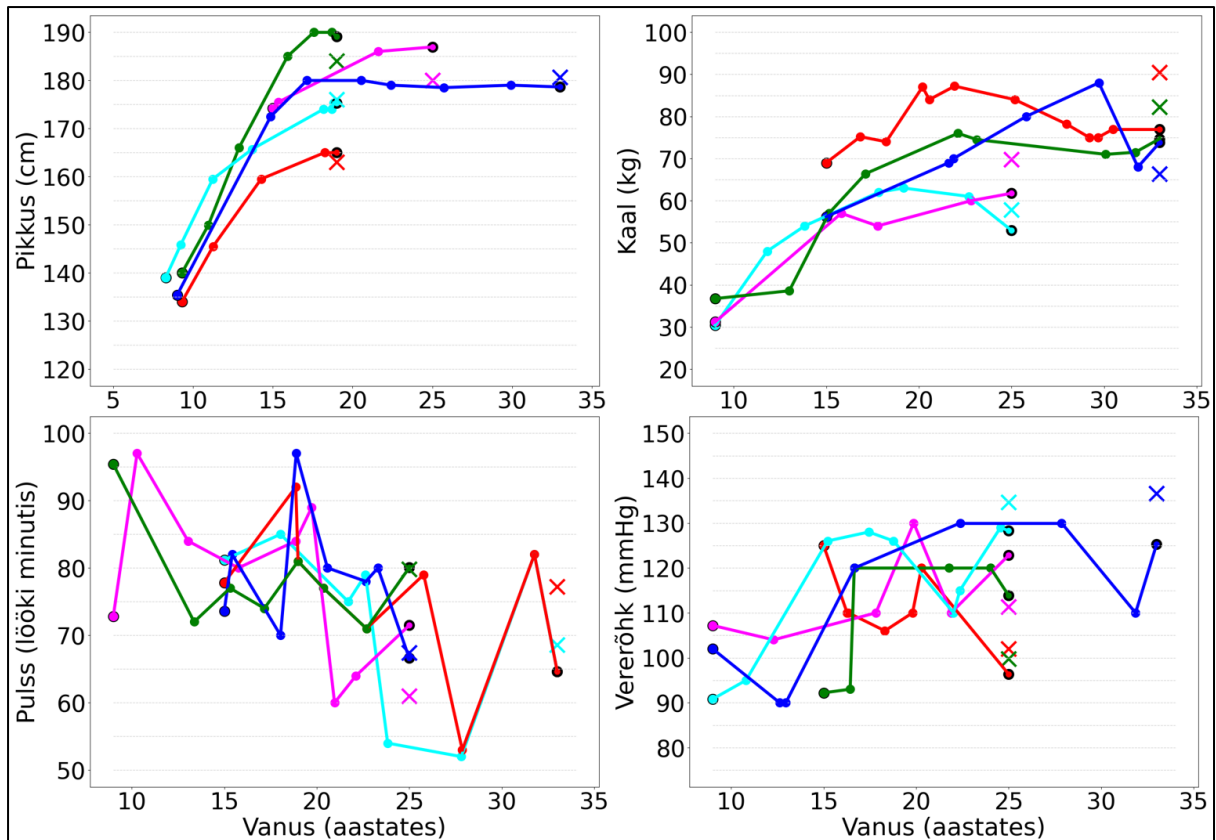
Tabel 2. RMSE väärtused eri mudelite ja testjuhtude puhul. Iga testjuhu täpseim tulemus on esile tõstetud.

Mudel	Treenimine	Testimine	„Kolme ennustuse keskmine“	„Andmestiku keskmine“
Pikkus (cm)	6.31	5.54	5.15	9.26
Kaal (kg)	19.07	19.40	16.06	18.92
Pulss (lööki minutis)	10.65	10.18	8.47	8.52
Vererõhk (mmHg)	13.78	23.54	20.16	14.90

Tabelis 2 on välja toodud sarnasuse mudeli treenimise ja testimise tulemused ning võrdluseks tehtud „andmestiku keskmine“ ja „kolme ennustuse keskmine“ mudeli tulemused vastavalt iga testjuhu jaoks.

3.3 Mudeli ennustuse protsess viie testjuhu näitel

Joonisel 7 on toodud iga nelja mudeli kohta viie juhuslikult testandmestikust võetud inimese algpunkt, tegelik lõpp-punkt, ennustatud lõpp-punkt ja ennustuse tegemiseks mudeli tehtud sammud. Iga värv tähistab erinevat ennustust. Vasakpoolne punkt on ennustuse lähtepunkt ning parempoolne on ennustatud mõõteväärtus. Tähega „x“ on toodud vastava isiku tegelik mõõteväärtus selles punktis.



Joonis 7. Sarnasuse mudeli ennustuse protsess nelja testjuhu viie juhusliku ennustuse näitel.

Iga testjuhu graafikul on näha vastava tervisenäitaja ennustamisprotsessi kulgemise trendi ja mustrit. Jooniselt saab ka välja lugeda, millistel testjuhtudel moodustub pikaajalisem kasvukõver ja millistel loogilist kulgu ei tundu olevat.

4. Arutelu

Selle bakalaureusetöö raames loodi sarnasuse otsimisel põhinev ennustusmudel, et ennustada inimeste pidevaid tervisenäitajaid ja demonstreeriti seda kehapiikkuse, kehakaalu, pulsi ja vererõhu näitel. Selle mudeli loomiseks kasutati RITA-MAITT andmestikku algandmetena ja ELIKTU uuringust pärinevaid andmeid treening- ja testandmetena. Sarnasuse hindamiseks kasutati mudelis skoorimist ning täpsuse hindamiseks RMSE-d.

Tabelist 2 ja mudeli ennustusprotsessi kirjeldavatest graafikutest (joonis 7) on näha, et sõltuvalt konkreetsest mõõdetavast tunnusest on mudeli täpsused küllaltki erinevad. Parima ennustustäpsusega oli pikkuse ennustusmudel, mille keskmiseks veaks oli treenimisel 6,3 ja testimisel 5,5 sentimeetrit. Kehakaalu ennustamise mudel oli aga treenimisel keskmiselt 19,1- ja testimisel 19,4-kilogrammise veaga. Kuigi need võivad tunduda suured eksimused, tuleb arvesse võtta, et treening- ja testandmestikud koosnesid peamiselt kasvueas lastest, kus alg- ja lõppmõõtmised erinesidki suurel määral – vahe oli pikkuse puhul keskmiselt 24,8 cm, kehakaalu puhul 33,1 kg ja ennustused tehti pika ajavahemiku peale, vahemikus 9–40 eluaastat.

Mudeli täpsusele konteksti andmiseks tehti töö raames lisaks sarnasuse otsimise mudelile ka „andmestiku keskmise“ ennustusmudel ja „kolme ennustuse keskmise“ mudel. „Andmestiku keskmise“ mudeli põhimõte on pakkuda kõikideks ennustusteks selle tunnuse andmestiku keskmist. Oluline on märkida, et päris rumala mudeliga tegemist ei ole, sest ennustusi ei pakuta juhuslikult, vaid mudel võtab arvesse ikkagi etteantud andmestiku keskmist ning seega piirab see eksimisvõimalust tunduvalt. „Kolme ennustuse keskmise“ ennustusmudeli töö põhimõte on sama sarnasuse otsimise mudeliga, kuid ühe sarnaseima inimese asemel otsitakse esimeses skoorimise punktis kolme. Lõplik ennustus on kõigi nende kolme haru tulemuse keskmine.

Tulemuste peatükis välja toodud RMSE-de tabelis (tabel 2) on ka toodud „andmestiku keskmise“ ja „kolme ennustuse keskmise“ mudeli tulemused iga nelja tunnuse kohta. Sellest tabelist on näha, et sarnasuse otsimise mudeli pikkuse ennustamise täpsus on „andmestiku keskmise“ mudeli omast 3,7 cm võrra täpsem. Kõikide teiste testjuhtude täpsused jäid aga „andmestiku keskmise“ mudeli täpsusest alla. Kui võtta aga ühe ennustuse asemel kolm ehk kasutada „kolme ennustuse keskmist“ mudelit, siis on näha, et kõik tulemused paranevad. Selle põhjuseks võib olla asjaolu, et kolme tulemuse keskmine vähendab riski, et ühel nendest kordadest satutakse ebatäpsetele lähteandmetele (vt täpsemalt allpool). „Kolme ennustuse keskmise“ mudeli tulemused ületavad ka „andmestiku keskmise“ mudeli täpsuse. Ainult

vererõhu testjuhu puhul jääb see kehvemaks, mille põhjuseks võib olla, et vererõhku on keeruline sellisel moel ennustada. Seda on näha ka tulemuste peatükis toodud ennustusprotsessi graafikutest (joonis 7), kus vererõhu kulgemisel ajas ei tundu olevat kindlat mustrit, kui aga pikkusel ja kaalul joonistub välja selge kasvukõver.

Et pikkuse mudeli treening- ja testandmed koosnesid enamasti kasvueas lastest, siis saab mingil määral pikkuse sarnasuse otsimise mudelit võrrelda ka kirjanduse ülevaates väljatoodud Mlakar jt loodud „kasvukõvera võrdlemise“ [19] meetodi tulemustega. Selleks ennustati pikkuse ennustusmudeliga eraldi pikkused kõigile isikutele ELIKTU andmestikust, kellel olid pikkused mõõdetud vähemalt 9- ja 25-aastaselt ning 15- ja 25-aastaselt. Nende kahe andmestiku ennustuste keskmise veaga saab võrrelda „kasvukõvera võrdlemise“ meetodi tulemusi, mis saadi mudelile lapse viimase pikkuse ette andmisega vastavalt 9- ja 15-aastaselt. „Kasvukõvera võrdlemise“ meetod eksis 9-aastaste puhul poiste ja tüdrukute kokkuvõttes keskmiselt 2,9 cm ja 15-aastaste puhul 1,7 cm. Sarnasuse otsimise meetod eksis 9-aastaste puhul keskmiselt 6,6 cm ja 15-aastaste puhul 3,8 cm. Mõlemal juhul ületab „kasvukõvera võrdlemise“ meetod täpsuselt mõningal määral sarnasuse otsimise meetodit. „Kolme ennustuse keskmine“ mudel eksis samadel andmetel vastavalt 5,0 cm ja 3,2 cm, mis on parem kui sarnasuse otsimise mudelil, kuid jääb „kasvukõvera võrdlemise“ meetodile alla. Samas tuleb arvestada, et „kasvukõvera võrdlemise“ mudel vajab lähteandmeteks vähemalt kahe mõõtepunkti andmeid, kuid selles bakalaureusetöös loodud mudelil piisab ühest. Silma jääb ka tendents, et mida hilisem on etteantud mõõtmine, seda täpsemalt suudab ennustusmudel soovitud vanuses pikkust ennustada.

Väljatöötatud mudelis kasutati sarnasuse hindamiseks nelja parameetrit: vanus, ennustatav tunnus ise (vastavalt siis pikkus, kaal, pulss ja vererõhk), võrreldava inimese viimase mõõtmise kaugus aastates praegusest punktist ja sugu. Tulemuste peatükis olevalt soojuskaardilt (joonis 6) on näha nende parameetrite kaalud. Kaalud on leitud juhusliku otsinguga hüperparameetrite häälestamise käigus. Mudeli sisendparameetrite väärtused ei ole küll sarnase jaotusega, kuid mudelite kaalusid võib järgnevalt siiski veidi analüüsida. Jooniselt 6 on näha, et vanuste vahe parameeter on pulsi ja vererõhu mudelite jaoks suurema kaaluga, kui on pikkuse ja kehakaalu mudelite jaoks. Vastupidiselt eelnevale on pulsi ja vererõhu mudelite jaoks patsiendi mõõtmiste ulatuvuse kaugus väiksema kaaluga kui pikkuse ja kehakaalu mudelite jaoks, mis tähendab, et viimased mainitud mudelid eelistavad ennustamisel neid patsiente, kelle kaudu saab teha pikemaajalise ennustuse. See on ka loogiline, sest, nagu eespool mainitud, moodustub

pikkuse ja kehakaalu ennustamisel selge kasvukõver ehk suure tõenäosusega liiguvad kõikide patsientide andmed ühes suunas. Sarnaselt patsiendi andmete ulatuvuse kaugusele on võrreldes pulsi ja vererõhu mudelitega kehakaalu ja pikkuse mudelite jaoks palju olulisem vastavate mõõtetulemuste vahe. Ka seda võib pidada loogiliseks, sest nagu kirjanduse ülevaates väljatoodud, on lapse hetkepikkus tulevase pikkuse ennustamiseks üks oluline tegur, kui aga pulsi ja vererõhu puhul võivad tähtsaks osutada teised tegurid. Sugu on kõikide testjuhu mudelite jaoks sarnaselt oluline, kuid kõige vähem olulisem on see siiski kehakaalu mudeli ja kõige olulisem pulsi mudeli jaoks. Soo parameetri kaalude tulemus on mõnevõrra üllatav, sest kaal on kõige suurem pulsi ennustamiseks, kuid just pulsi ELIKTU andmete graafikul (joonis 4) ei näi poiste ja tüdrukute vahel olevat tegelikult suurt erinevust.

Loodud mudeli üks puudus on selle suur sõltuvus üksikutest mõõtetulemustest lähteandmestikus, mis võivad olla vigased. Kasutatud RITA-MAITT uuringu andmetes on suur osa mõõtmistest eraldatud tekstilistest algandmetest, kus lisaks juba sisestusvigadele võis tekkida vigu ka mõõdetud tunnuse, väärtuse, ühiku ja kuupäeva eraldamisel. Isegi kui eemaldada andmetest äärmuslikud väärtused, on ebareaalselt suuri kõikumisi ilma meditsiinilisi teadmisi omamata andmetes raske tuvastada. Nende vigade tõttu võib juhtuda, et mudel jõuab isikuni, kelle andmetes on viga. Need vead võivad omakorda mõjutada järgmist sarnasuse otsimist, põhjustada suure ennustusvea ning mõjutada parameetrite häälestamist ja RMSE tulemust. Sellise olukorra vastu võib aidata „kolme ennustuse keskmise“ meetodi kasutamine. „Kolme ennustuse keskmise“ meetodi edukus võib viidata ka sellele, et täpsust on võimalik veelgi parandada, võttes kolme ennustuse keskmise asemel näiteks viie, seitsme või lausa kümne ennustuse keskmise.

Kaugeleulatuvate ennustuste mudeli treenimiseks on vaja esinduslikke treening- ja testandmeid. Paraku on selliseid andmeid väga vähe. Töös kasutatud treening- ja testandmed pärinevad uuringust ELIKTU. See andmestik koosnes inimeste mõõtmistest eluaastatel 8–34 ning inimestel mõõdeti samal ajal nii pikkust, kaalu, pulssi kui ka vererõhku, mistõttu sobis see andmestik siia töösse hästi. Kuna kõiki inimesi ELIKTU andmestikus mõõdeti samas vanuses, võib mudel olla kallutatud ennustama just selles vanusevahemikus ja vanusevahelgi inimesi. Tuleb ka arvestada, et ELIKTU ei sisalda isikuid, kelle vanus on rohkem kui 34 eluaastat. Seetõttu võib treening- ja testandmestikku lugeda küllaltki kallutatuks. Sobivama andmestiku saamiseks tuleks andmestikku täiendada nii, et seal oleksid inimeste vanused ühtlasemalt jaotunud ning vahed algse vanuse ja lõppvanuse vahel varieeruksid rohkem.

Üks võimalik mudeli edasiarendus on parameetrite valiku suurendamine vastavalt konkreetset tunnust kõige enam mõjutatavatele teguritele. Võib oletada, et see võimaldaks parandada tervisenäitajate, millel on veel teisi tähtsaid mõjutavaid tegureid peale vanuse ja soo, ennustamise täpsust. Samuti saaks kasutada sama tunnuse varasemaid mõõtmisi. Hetkel vastavate testjuhtude mudelites kasutusel olevad tegurid on kõik olemas ühes ja samas andmepunktis. Sellisel lähenemisel võib tekkida probleem aga sellega, et kõik mudelisse juurdevõetud uued tegurid ei pruugi olla kirja pandud või mõõdetud üksteisega samas punktis ehk samal ajahetkel. Üks selle probleemi lahendus on võtta ainult need punktid andmestikust, kus on olemas kõik vajalikud tegurid, aga see võib liigselt vähendada andmestiku hulka. Samuti võib lahendus olla see, kui punktis puuduolevad tegurid võtta iga inimese lähimast järgmisest või eelmisest punktist, mis pika ajavahe peale võib aga tekitada suuri vigu.

Täiustada saaks ka parameetrite kaalude optimeerimist. Hetkel kasutab mudel juhuslikku kaalude valimist, mille puhul ei pruugi mudel optimaalseid kaalusid leida ning kaalude suurus võib igal korral ka varieeruda. Optimaalsete parameetrite leidmiseks tuleks proovida kasutada efektiivsemaid meetodeid. Optimeerida saaks ka skooride arvutamist, mida praegu arvutatakse andmestikus olevate inimeste kõikidele andmepunktidele, mis võtab palju aega. Seda saaks aga mingitel tingimustel piirata, et mudel ei arvutaks iga kord igale andmepunktile skoori, vaid teeks seda ainult potentsiaalselt sobivatele andmepunktidele, näiteks pikkuse puhul ainult viie sentimeetri raadiuses olevatele punktidele.

Kokkuvõte

Bakalaureusetöö eesmärk oli uurida, kas ja kuidas on võimalik luua pidevate tervisetunnuste prognoosimiseks ennustusmudel, mis suudab teha kaugeleulatuvaid prognoose suhteliselt lühikesest ajavahemikust pärit andmete põhjal. Autorile teadaolevalt ei ole sarnast meetodit varem kasutatud.

Töös näidati, et selline lähenemine on võimalik ning loodi sarnasuse otsimisel põhinev ennustusmudel nelja testjuhu jaoks: kehapikkus, kehakaal, pulss ja süstoolne vererõhk. Samal põhimõttel loodi ka „kolme ennustuse keskmise“ ennustusmudel. Kõikide testjuhtude tulemusi võrreldi „andmestiku keskmise“ ennustusmudeliga ning pikkuse testjuhu mudelit võrreldi ka kirjanduses varem avaldatud „kasvukõvera võrdlemise“ meetodiga.

Kui kehapikkuse ennustamine andis küllaltki täpseid tulemusi, siis teiste testjuhtude mudelid jäävad „andmestiku keskmise“ ennustusmudelist ebatäpsemaks. „Kolme ennustuse keskmise“ ennustusmudelit kasutades paranevad kõikide testjuhtude tulemused ning need ületavad ka „andmestiku keskmise“ mudeli täpsust (v.a vererõhk). Töös on kirjeldatud tulemuste võimalikke põhjuseid ja lahendusvariante.

Kuigi kõikide pidevate tervisenäitajate kaugeleulatuvaiks ennustamiseks ei pruugi see meetod sobida, annab mõnede näitajate ennustamine küllaltki täpseid tulemusi ja seetõttu väärrib meetod edasi uurimist ja täiustamist. Edaspidi, kui selle meetodiga tehtud mudelid lähevad keerukamaks ja täpsemaks, saab mudelit tervisevaldkonnas praktiliselt rakendada näiteks kehakaalu ennustamisel, et juba varakult sekkuda ja tänu sellele hoida ära ülekaalu ja sellest tekkivaid terviseprobleeme.

Tänu sõnad

See bakalaureusetöö on läbi viidud uuringute RITA1/02-96 ja PRG1844 raames.

Viidatud kirjandus

- [1] Kernbach J. M., Staartjes V. E. Machine learning-based clinical prediction modeling Part I: Introduction and general principles. *Machine learning-based clinical prediction modeling A practical guide for clinicians*, 2020, Vol. 2.1, p. 2-12.
- [2] James M. Predictive Modeling: History, Types, Applications. *Investopedia*, 2023. <https://www.investopedia.com/terms/p/predictive-modeling.asp#toc-history-of-predictive-modeling> (05.04.2024).
- [3] Kaalep H.-J., Koit M. Kuidas masin tõlgib. Keel ja Kirjandus, 2010, nr 10, lk 726–738.
- [4] Chen L. Overview of clinical prediction models. *Annals of Translational Medicine*, 2020, Vol. 8, No. 4, p. 71. <https://doi.org/10.21037/atm.2019.11.121>
- [5] Burnham J. P., Lu C., Yaeger L. H., Bailey T. C., Kollef M. H. Using wearable technology to predict health outcomes: a literature review. *Journal of American Medical Informatics Association*, 2018, Vol. 25, Iss. 9, p. 1221-1227. <https://doi.org/10.1093/jamia/ocy082>
- [6] Yasar K., Wigmore I. Definition wearable technology. *TechTarget*, 2023. <https://www.techtarget.com/searchmobilecomputing/definition/wearable-technology> (03.05.2024).
- [7] Jenkins D. A., Sperrin M., Martin G. P., Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and Prognostic Research*, 2018, Vol. 2, No. 23. <https://doi.org/10.1186/s41512-018-0045-2>
- [8] Lapp L., Roper M., Kavanagh K., Bouamrane M.-M., Schraag S. Dynamic Prediction of Patient Outcomes in the Intensive Care Unit: A Scoping Review of the State-of-the-Art. *Journal of Intensive Care Medicine*, 2023, Vol. 38, Iss. 7, p. 575-591. <https://doi.org/10.1177/08850666231166349>
- [9] A transformer method that predicts human lives from sequences of life events. *Nature Computational Science*, 2023, Vol. 4, p. 7-8. <https://doi.org/10.1038/s43588-023-00586-0>
- [10] Mello-Klein C. A new AI model can predict human lifespan, researchers say. They want to make sure it's used for good. *Northeastern Global News*, 2023. <https://news.northeastern.edu/2023/12/19/predictive-ai-human-lifespan-model/> (03.04.2024).
- [11] Iannelli V. 3 Methods for Predicting Your Child's Height. *verywell family*, 2020. <https://www.verywellfamily.com/kids-height-predictors-3879394> (04.04.2024).
- [12] American Academy of Pediatrics. Predicting a Child's Adult Height. 2016. <https://www.healthychildren.org/English/health-issues/conditions/Glands-Growth-Disorders/Pages/Predicting-a-Childs-Adult-Height.aspx> (04.04.2024).
- [13] Zeevi D., Yehuda A. B., Zangen D., Kruglyak L. Accurate prediction of children's target height from their mid-parental height. *medRxiv*, 2022. <https://doi.org/10.1101/2022.10.31.22281712>

- [14] Height Potential: A Child's Target Height Based on Midparental Height. s.a. <https://www.merckmanuals.com/en-ca/professional/multimedia/clinical-calculator/height-potential-a-childs-target-height-based-on-midparental-height> (04.04.2024).
- [15] Georgiev GZ. Height Calculator. s.a. (04.04.2024).
- [16] World Health Organization. Height-for-age Girls. 2007. [https://cdn.who.int/media/docs/default-source/child-growth/growth-reference-5-19-years/height-for-age-\(5-19-years\)/cht-hfa-girls-perc-5-19years.pdf?sfvrsn=f90a33cf_4](https://cdn.who.int/media/docs/default-source/child-growth/growth-reference-5-19-years/height-for-age-(5-19-years)/cht-hfa-girls-perc-5-19years.pdf?sfvrsn=f90a33cf_4) (20.04.2024).
- [17] Cordeiro J. R., Postolache O., Ferreira J. C. Child's Target Height Prediction Evolution. *Applied Sciences*, 2019, Vol. 9, Iss. 24, p. 5447. <https://doi.org/10.3390/app9245447>
- [18] Shmoish M., German A., Devir N., Hecht A., Butler G., Niklasson A., Albertsson-Wikland K., Hochberg Z. Prediction of Adult Height by Machine Learning Technique. *The Journal of Clinical Endocrinology & Metabolism*, 2021, Vol. 108, Iss. 7, p. 2700-2710. <https://doi.org/10.1210/clinem/dgab093>
- [19] Mlakar M., Gradišek A., Luštrek M., Jurak G., Sorić M., Leskošek B., Starc G. Adult height prediction using the growth curve comparison method. *PLoS ONE*, 2023, Vol. 18, Iss. 2. <https://doi.org/10.1371/journal.pone.0281960>
- [20] University of Arkansas for Medical Sciences. Do people shrink as they age? *UAMS Health*, 2019. <https://uamshealth.com/medical-myths/do-people-shrink-as-they-age/> (11.04.2024).
- [21] Weingarten J. What Are the Mechanisms and Effects of Age-Related Shortening of the Spine? *The Science Journal of the Lander College of Arts and Sciences*, 2018, Vol. 12, No. 1. <https://touro scholar.touro.edu/sjlcas/vol12/iss1/8/> (11.04.2024).
- [22] Sorkin J. D., Muller D. C., Andres R. Longitudinal Change in Height of Men and Women: Implications for Interpretation of the Body Mass Index: The Baltimore Longitudinal Study of Aging. *American Journal of Epidemiology*, 1999, Vol. 150, Iss. 9, p. 969-977. <https://doi.org/10.1093/oxfordjournals.aje.a010106>
- [23] WebMD Editorial Contributors. How Much Do You Shrink As You Age? WebMD, 2023. <https://www.webmd.com/healthy-aging/how-much-do-you-shrink-as-you-age> (12.04.2024).
- [24] Oja M., Tamm S., Mooses K., Pajusalu M., Talvik H.-A., Ott A., Laht M., Malk M., Lõo M., Holm J., Haug M., Šuvalov H., Särg D., Vilo J., Laur S., Kolde R., Resiberg S. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *JAMIA Open*, 2023, Vol. 6, Iss. 4. <https://doi.org/10.1093/jamiaopen/ooad100>
- [25] Talvik H.-A., Oja M., Tamm S., Mooses K., Särg D., Lõo M., Siimon Õ. R., Šuvalov H., Kolde R., Vilo J., Reisberg S., Laur S. Repeatable Process for Extracting Health Data from HL7 Cda Documents. 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4776237 (04.05.2024).

[26] ELIKTU. LASTEUURING. s.a. <https://www.ecpbhs.ee/andmekogumine/lasteuuring/>
(24.04.2024).

LISA 1: Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Belinda Lepmets,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Pidevate tervisenäitajate kaugeleulatav ennustamine“, mille juhendaja on Sulev Reisberg, 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 Commons'i 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.

Belinda Lepmets

14.05.2024