TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Egle Saks

Eesti alaliste elanike määramine kasutades masinõppe meetodeid

Magistritöö (15 EAP)

Juhendaja(d): Terje Trasberg, PhD
Raivo Kolde, PhD

Tartu 2023

Eesti alaliste elanike määramine kasutades masinõppe meetodeid

Lühikokkuvõte:

Riiklikul statistikal on oluline roll levitada ühiskonna kohta teadmisi ja fakte, mis võimaldaksid teha informeeritud otsuseid. Üks olulisemaid riikliku statistika levitavaid teadmisi on info rahvastiku kohta ning selle keskmes on info rahvaarvu kohta. Järjest kiiremini muutuvas maailmas vananeb informatsioon kiiremini kui varem ning seega oodatakse ka rahvastikustatistikat kiiremini ja tihemini. Euroopa Komisjon valmistab juba ette määrust, millega tuleks alaliste elanike arvu riigis avaldada kaks korda aastas. Praegu pannakse Eestis alalise elanikkonna kogum kokku kasutades 18 erinevat registrit, mis muudab tihemini avaldamise keeruliseks.

Selle magistritöö eesmärk on uurida, millised andmed on residentsuse määramiseks kõige olulisemad ja kuidas saavad elanikkonna määramisega vähendatud andmete kontekstis hakkama masinõppe mudelid. Töö eesmärgi täitmiseks on kasutatud Eesti Statistikaameti poolt kättesaadavaks tehtud andmeid. Andmetel rakendatakse peakomponentide analüüsi ning testitakse viit erinevat masinõppe mudelit. Tulemused näitavad, et vähendatud andmestik toimib üsna võrdväärselt algse andmestikuga ning residentsuse tuvastamiseks võib piisata ka väiksemast hulgast registritest. Masinõppe meetoditest toimivad kõige paremini otsustusmets ja XGBoost.

Võtmesõnad: Registrid, alaline elanik, masinõpe, rahvastik, statistika

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Determining Estonian Usual Residents Using Machine Learning Methods

Abstract:

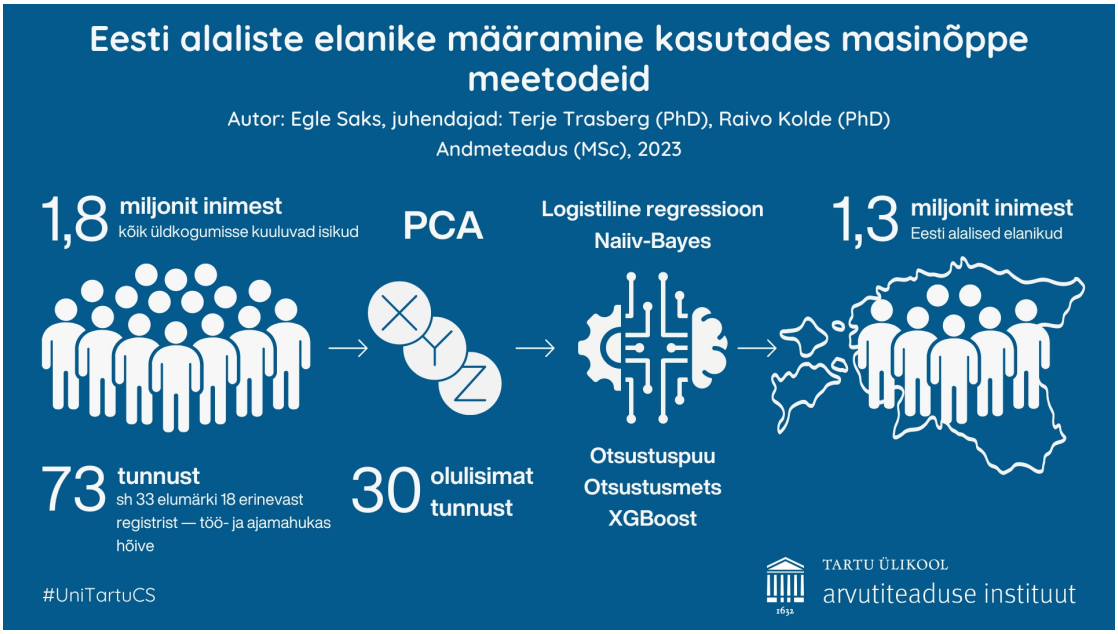
Official statistics play an important role in disseminating knowledge and facts about society, enabling informed decision-making. One of the most important pieces of information disseminated by official statistics is information about the population, with the population size being at the center of this. In an increasingly fast-paced world, information becomes outdated more quickly than ever before, meaning that population statistics are expected more frequently and regularly. The European Commission is preparing a regulation to require the publication of usually resident population twice a year. However, in Estonia, compiling usually resident population using 18 different registers makes more frequent publication challenging.

The aim of this master's thesis is to investigate which data is most important for determining residency and how machine learning models can handle population determination in the context of reduced data. Data used for the purpose of this study is made available by the Statistics Estonia. Principal component analysis is applied to the data, and five different machine learning models are tested. The results show that the reduced dataset performs quite similarly to the original dataset, and a smaller set of registers may be sufficient for determining residency. Among the machine learning methods tested, Random Forest and XGBoost perform the best.

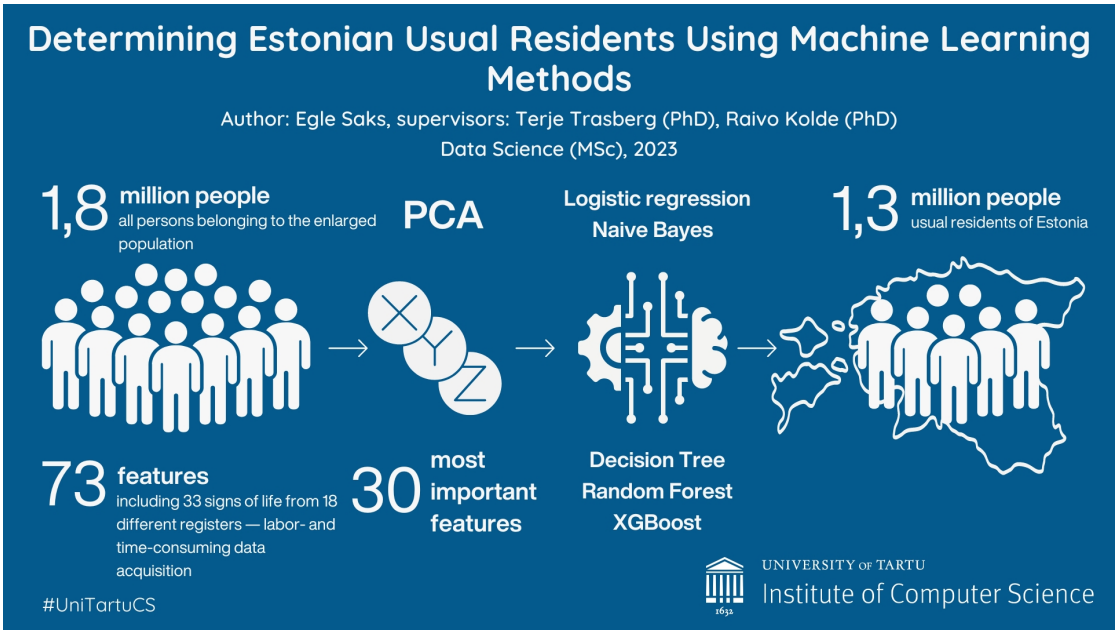
Keywords: Registers, usual resident, machine learning, population, statistics

CERCS: P160 Statistics, operation research, programming, actuarial mathematics

Visuaalne kokkuvõte



Visual abstract



Sisukord

Sissejuhatus

Vajadust pidada arvet inimkapitali üle on peetud oluliseks ühiskonna algusajast. Tänapäeval mõistetakse seda kui rahvastikustatistikat, mille kogumisega tegelevad peamiselt riiklikud statistikaametid. Ametliku statistika põhjal teevad otsuseid paljud isikud ja asutused ning statistika abil saab ühiskond juurde uusi teadmisi. Usaldusväärne statistika on eriti oluline 21. sajandil, pakkudes kaitset valeinfo levimise eest.

Statistika kvaliteet ja kasulikkus sõltub mitmetest teguritest. Näiteks OECD (2017) peab statistika kvaliteedi puhul oluliseks kaheksat tegurit: asjakohasus, täpsus, usaldusväärsus, ajakohasus, juurdepääsetavus, tõlgendatavus ja sidusus. Niisiis on oluline, et statistiline teave oleks väärtuslik ning aktuaalne, et oleks võimalik sellest lähtuvalt tegutseda ja teha otsuseid. Ka institutsioonid ise on huvitatud statistika ajakohasusest. Näiteks on Euroopa Komisjon ette valmistamas rahvastiku ja eluruumide statistika kohta uut määrust, kus muuhulgas on üks kavadatavatest muutustest see, et rahvaarvu ehk alaliste elanike arvu tuleks hakata esitama kaks korda aastas senise ühe korra asemel. Määrust soovitakse kehtestada 2026. aastaks.

Töö probleem seisneb selles, et Eesti elanike määramine praegusel kujul vajab palju sisendandmeid, mille hõive võtab aega ning nõuab kokkuvõttes palju inimressurssi andmete vastuvõtmise, andmebaasi laadimise, kontrollimise jms näol. Kuna sisendandmeid on palju, on ka tulemuse seletatavus keerulisem, kui väiksema hulga sisendandmete kasutamisel. Selle magistritöö eesmärk on uurida, millised andmed on residentsuse määramiseks kõige olulisemad ja kuidas saavad elanikkonna määramisega vähendatud andmete kontekstis hakkama masinõppe mudelid. Töö uurimisküsimused on:

1. Millised tunnused praegu olemasolevatest andmetest on kõige vajalikumad, et määrata residentsust?
2. Kuidas toimivad vähendatud andmestikul masinõppe algoritmid?

Magistritöö koosneb neljast peatükist: teoreetiline taust, metoodika, tulemused ja arutelu. Teoreetilises osas antakse ülevaade põhimõistetest ja -kontseptsioonidest nagu rahvastikustatistika ja alaline elanikkond. Lisaks kirjeldatakse praegust Eesti

alaliste elanike määramise metoodikat ning Euroopa Liidu riikide praktikaid alaliste elanike määramiseks. Metoodika peatükis antakse ülevaade rakenduslikus osas kasutatud andmetest ja metoodilistest valikutest. Tulemuste peatükis kirjeldatakse eksperimentide käigus saadud tulemusi ja hinnatakse neid. Arutelu peatükis võetakse tulemused kokku ning antakse edasisi uurimissoovitusi.

1 Teoreetiline taust

Selles peatükis antakse ülevaade uurimistöö teemaga seotud mõistetest ja kontseptsioonidest. Tuuakse välja rahvastikustatistikas olulised definitsioonid ning statistika avaldamise õiguslik alus. Kirjeldatakse registripõhist statistikat ning kuidas see on kasutusel Eestis residentsuse indeksi puhul. Samuti antakse ülevaade alaliste elanike kokkupaneku praktikatest teistes Euroopa riikides.

1.1 Teemapüstitus ja mõisted

Riiklik statistika on statistika, mida koostavad ja avaldavad valitsusasutused ja muud avalik-õiguslikud asutused. Kõige suurem asutus, mis ametlikku statistkat riigi kohta kogub ja avaldab on tavaliselt riiklik statistikaamet. Eestis on riikliku statistika tegemine reguleeritud riikliku statistika seadusega. Riikliku statistika seaduse kohaselt on riikliku statistika tegijad Eestis Statistikaamet, mis teeb riiklikku statistikat ning teostab järelevalvet ning Eesti Pank, kes teeb ja avaldab raha-, finants- ja maksebilansistatistikat. (Rahandusministeerium, 2022)

Riiklik statistika on määratletud kui *kvantitatiivne, kvalitatiivne, kokkuvõtlik ja üldistav teave, mis iseloomustab massnähtust vaatlusaluses kogumis ja mis saadakse riikliku statistika programmi või programmivälise statistikatöö raames andmete statistilise töötlemise tulemusena* (RT, 2022). Riikliku statistika seadus annab statistikaametile õigusliku aluse andmejagamisteenusele ehk õiguse saada andmeid riigi- ja kohaliku omavalitsuse asutuselt või avalik-õiguslikult juriidiliselt isikult (RT, 2022). Lisaks Eesti seadustele kohalduvad Eesti riiklikule statistikale ka Euroopa Liidu õigusnormid.

Riiklik statistika jaguneb paljudeks valdkondadeks, mille alla kuuluvad muuhulgas majandus-, keskkonna-, sotsiaal- ja rahvastikustatistika. Neist viimane on selle töö fookuses.

1.1.1 Rahvastikustatistika

Rahvastikustatistika näitab rahvastiku suurust (rahvaarvu ehk alaliste elanike arvu) ja kajastab rahvastiku muutuseid, nagu sünnid, surmad, sisseränne ja väljaränne. Rahvastikustatistikas edastatavad andmed on reguleeritud Euroopa Parlamendi ja nõukogu määrustega 862/2007 ja 1260/2013 ja nende rakendusmäärustega, kuid hetkel on Euroopa Komisjon ette valmistamas uut rahvastikustatistika määrust. Uues määruses soovitakse alaliste elanike arvu soo ja vanuse lõikes senise ühe korra asemel kaks korda aastas — 31. detsembri ja 30. juuni seisuga. (Euroopa Komisjon, 2021)

Uus määrus soovitakse jõustada alates 2026. aastast. Euroopa Komisjon põhjendab muudatuse vajadust sellega, et tulenevalt demograafilistest muutustest ja rändesuundumustest on statistika rahvastiku kohta muutunud viimase kümne aastaga poliitiliselt ja ühiskondlikult olulisemaks kui varem ning seega on oluline, et statistika oleks ajakohasem ja sagedasem. (Euroopa Komisjon, 2021)

Kuigi Euroopa riikides on rahvaarvu arvutamise metoodikad erinevad ning mitte kõik riigid ei kasuta registreid, tähendab sagedasem avaldamine ka registripõhist statistikat ja uusi metoodikaid. Praegu Eesti alaliste elanike määramiseks kasutatav residentsuse indeks hõlmab 18 erinevat registrit, mille andmete hõive ja kasutamine on töö- ja ajamahukas. Andmed tuleb laadida andmebaasi, kontrollida ja küsimuste või probleemide korral suhelda andmekoguga. Uuest Euroopa Komisjoni määrusest tuleneb selle töö eesmärk. Magistritöö eesmärk on uurida, millised andmed on residentsuse määramiseks kõige olulisemad ja kuidas saavad elanikkonna määramisega vähendatud andmete kontekstis hakkama masinõppe mudelid. See töö panustab sellesse, et Eesti alaliste elanike määramine saaks toimuda tulevikus kiiremini ja efektiivsemalt ning alaliste elanike arvu saaks tulevikus avaldada ka aasta keskel.

1.1.2 Alalise elaniku definitsioon

Vastavalt määrusele 1260/2013 edastavad liikmesriigid Euroopa Komisjonile (Eurostatile) igal aastal andmed oma alalise elanikkonna kohta vaatlusajal (31. detsembri südaöö). Vastavalt määrusele 1260/2013 on alaline elanik isik, kes:

- on elanud oma alalises elukohas pidevalt vähemalt 12 kuud enne vaatlusaega,
- või on saabunud oma alalisse elukohta vaatlusajale eelnenud 12 kuu jooksul kavatsusega elada seal vähemalt üks aasta. (ELT, 2013)

Alalisse elanikkonda kuuluvad kõik isikud, kelle alaline elukoht on vaatlusajal liikmesriigis. Edastatavad andmed hõlmavad elanikkonna jagunemist vanuse, soo ja elukoha piirkonna lõikes. (ELT, 2013)

Praegu pannakse Eestis alalise elanikkonna kogum kokku kasutades 18 erinevat riiklikku andmekogu ja registrit (Statistikaamet, 2022). Alaliste elanike kogumi kokkupanekul registripõhiselt on kesksel kohal katvuse küsimus, sest eesmärgiks on saavutada sihtrühmale võimalikult lähedane tulemus. Katvus (ingl k *coverage*) iseloomustab statistika piire ehk ulatust. Tavaliselt on vajalik määratleda kolm dimensiooni: ruumiline mõõde, ajaline mõõde ja kontseptuaalne mõõde. Ruumiliseks mõõtmeks selle töö kontekstis on Eesti riik, ajaliseks mõõtmeks 30. juuni/31. detsembri südaöö ning kontseptuaalseks mõõtmeks on alalised elanikud. Kaetuse seisukohalt on oluline jälgida kahte aspekti. Esiteks, kas administratiivsed andmed katavad soovitud sihtrühma. Teiseks, millised kaetuse probleemid tulenevad administratiivses andmeallikas puuduolevast infost. (CROS, 2019b)

Sihtrühm on eesmärgiks olev populatsioon, mille kohta teavet statistika kujul edastatakse. Sihtrühma piiritlemiseks registripõhises statistikas kasutatakse administratiivseid allikaid. Ideaalse katvuse puhul on kogu sihtrühm kaetud. Sihtrühmaks selles töös on Eesti alalised elanikud. Sihtrühm on esindatud rahvastikuregistris ja teistes registrites. (CROS, 2019c)

1.1.3 Alaliste elanike määramine registrite põhjal

Administratiivne register on register, mida kasutatakse administratiivsetel ehk halduslikel eesmärkidel. Administratiivne register on administratiivse andmeallika üks tüüp. Administratiivsest registrist saab statistiline register pärast statistilist töötlust, mis muudab selle sobivaks statistiliseks otstarbeks. Administratiivne register koondab infot sihtrühma kohta ning selle ulatus on tavaliselt määratletud seadusega. (CROS, 2019a) Administratiivsed süsteemid mäletavad kõike — need süsteemid katavad

kogu populatsiooni ja igal ajal. Inimesed rahvastikuregistris, ettevõtted, asutused ja organisatsioonid äriregistris on kõik osalejad ühiskonnas. Nende tegevus loob suurt hulka andmeid paljudes administratiivsetes ametiasutustes, mis tegelevad maksudega, heaoluteenustega nagu haridus, tervishoid ja sotsiaalhoolekanne. (A. Wallgren ja B. Wallgren, 2022)

Administratiivseid registreid kasutatakse klassikaliste küsitluste asemel rahvastikustatistikas järjest rohkem. Selleks on kaks põhjust: kulude vähendamine ja kvaliteedi tõstmine. Kvaliteedi poole pealt on näha, et klassikaliste küsitluste puhul toimub pidev vastajamäärade langus, mis tekitab küsimusi tulemuste kvaliteedis. Lisaks on näost-näkku küsitluste läbiviimine kallis ning poliitiline surve sunnib küsitluste hulka vähendama, et töötada tõhusamalt. Peale selle, et registrite kasutamine on statistikaametite jaoks tõhus viis väärtusliku teabe kogumiseks, vähendab see oluliselt inimeste ja majapidamiste aruandluskoormust. Administratiivsed registrid võimaldavad ka paindlikkust uute ja ennenägematute statistikavajaduste jaoks. (Bakker, Rooijen ja Toor, 2014; A. Wallgren ja B. Wallgren, 2022)

Kuid ka administratiivsete registrite puhul on oluline kvaliteeti kontrollida ning katvus on üks peamisi kvaliteedinäitajaid. Erinevus tegeliku ja registreeritud rahvaarvu vahel on registri katvusviga. Surmad ja väljaränne, mida ei ole teada antud enne kogumi kokkupanekut põhjustab ülekaetuse, samal ajal kui sünnid ja sisseränne, mida ei ole registreeritud toovad kaasa alakaetuse. Administratiivsete allikate põhjal tehtavas statistikas on seega vajalik allikate kombineerimine, sest allikaid kombineerides on suurem tõenäosus saada ideaalsele sihtrühmale võimalikult lähedane tulemus. Kaetuseprobleemi statistilises registris saab vähendada kasutades kõiki asjakohaseid allikaid. Selleks tuleb kõigepealt kaardistada kõik allikad. (A. Wallgren ja B. Wallgren, 2022)

Lisaks on oluline eristada kahte erinevat tüüpi registreid ja andmeid: a) universaalsed registrid, b) osalised registrid ehk registrid, mis kajastavad inimese tegevust. Universaalne register püüab sisaldada kogu populatsiooni, samas kui osaline register on limiteeritud teatud grupile, nt koolilapsed, valijad, sotsiaaltoetuste saajad. (Poston, 2019) Alalise elanikkonna määramiseks on vaja kasutada ka teist tüüpi registreid, mis kajastavad inimese aktiivsust. Näiteks inimesi, kes on rahvastikuregistris ja tööregistris ja saavad

sotsiaaltoetusi on vähem kui inimesi ainult rahvastikuregistris. Samas võib olla ka isikuid, kes on aktiivsust väljendatavates registrites, kuid mitte rahvastikuregistris ehk on rahvastikuregistris alakaetud. (A. Wallgren ja B. Wallgren, 2022)

Registriaktiivsust saab kasutada alalise elanikkonna piiritlemiseks järgnevalt:

1. Inimesed, kes on rahvastikuregistris ja omavad aktiivsust/elumärke teistes registrites. Need inimesed ilmselt kuuluvad alalisse elanikkonda.

2. Kaheldav kategooria ehk inimesed, kes on rahvastikuregistris aga kellel ei ole aktiivsust/elumärke teistest registrites. Kaheldava kategooria saab defineerida erinevat moodi, näiteks inimesed, kellel ei ole viimase kahe aasta jooksul ühtegi elumärki indikaatorregistrites. Need inimesed tõenäoliselt ei kuulu alalisse elanikkonda ehk on ülekaetud. Nende puhul on valik kolme reegli vahel: kõik sisse arvata, kõik välja arvata või sisse arvestada osad vastavalt mingile kriteeriumile.

3. Inimesed, kes ei ole rahvastikuregistris, aga on aktiivsed teistes registrites. Need inimesed kuuluvad tõenäoliselt alalise elanikkonna hulka, nad on alakaetud. Lähtudes sisulistest kaalutlustest tuleks defineerida, mis kategooriad sisse arvestada. Reegel peaks defineerima alalised elanikud ja mingil viisil ajutised elanikud, kes õpivad või töötavad riigis kui rahvastiku liikmed. (A. Wallgren ja B. Wallgren, 2022)

Lisaks aktiivsust väljendatavate registrite kasutamisele on oluline registreid prioritiseerida ning selleks on kõige levinum meetod kaalude kasutamine. Kaalusid saab kasutada, et korrigeerida arvestuslikku ülekaetust. Enne korrigeerimist on kõik kaalud võrdsed ühega. Pärast korrigeerimist, on kategooriatele, mis on ülekaetud, kaalud väiksemad kui üks. Kõik teised statistilised tööd, mis kasutavad rahvastikuregistrit, peaksid kasutama samuti neid kaale. Seega, kogu avaldatav statistika saab olema järjepidevalt korrigeeritud arvatava ülekaetuse efekti osas. (A. Wallgren ja B. Wallgren, 2022)

Registripõhise statistika puhul on tähtis ka ajaline määratletus, et oleks võimalik leida seisund erinevatel hetkedel. Aeg on meie igapäevaelu tajumises ja meid ümbritseva maailma sündmuste interpreteerimises ülitähtis komponent. Info aja kohta peab seega

olema oluline osa ka registri sisust. Administratiivsetes registrites on erinevat tüüpi ajaviited. Tuleb eristada kuupäeva, millal reaalses maailmas sündmus juhtus ja millal see sündmus mõõdeti või registreeriti. (Euroopa Komisjon, 1995) Rahvastikuregister põhineb kõigil demograafilistel sündmustel nagu sünnid, surmad ja ränne. Kõigil neil sündmustel on viide aja kohta: päev, millal sündmus toimus ja päev, millal see sündmus registreeriti registrisse.

Rahvastikuregistris, aga ka muudes registrites on seega neli erinevat tunnust, mis viitavad ajale:

1. alguskuupäev: aeg millal sündmus toimus, näiteks isik koodiga 7048 kolis eluruumi 339 1. oktoobril 2010.

2. registreerimise kuupäev: aeg, millal rahvastikuregistrit uuendati informatsiooniga selle sündmuse kohta. 5 novembril 2010. aastal uuendati registri isikut 7048 puudutavat infot.

3. lõppkuupäev: aeg millal järgmine seda inimest puudutav sündmus toimus, näiteks surm, väljaränne. Endiselt aktiivsetel vaatlustel seda väärtust ei ole.

4. deregistreerimise kuupäev: aeg, millal vaatlus muutus registris aktiivsest mitteaktiivseks. (A. Wallgren ja B. Wallgren, 2020)

Niisiis eeldab ülekaetuse ja alakaetuse vähendamine *elumärke* ehk registriaktiivsust kajastavaid allikaid, mis on Eestis residentsuse indeksis ka kasutusel. Siiski on ressursimahukas kasutada tulevikus tihedamaks avaldamiseks 18 erinevat registrit ja 33 elumärki nagu praegu. Samuti võib allikate paljusus mingist hetkest suurendada ka ülekaetust ehk vähendada meetodi täpsust. Masinõppes kasutatakse selle fenomeni kohta mõistet *curse of dimensionality*. Väiksem tunnuste hulk lihtsustab andmete visualiseerimist ja tulemuse mõistetavust, suurendab mudeli kiirust, kasutatavust ja täpsust. Selleks, et aasta keskel toimuv rahvastiku avaldamine oleks ajaliselt efektiivne ja ressursitõhus, et ei oleks vaja 18 registrilt küsida andmeid mitu korda aastas, on vajalik leida, millised registrid ja elumärgid on piisavad, et täpne alaliste elanike kogum kokku panna.

1.2 Alaliste elanike määramise metoodika Eestis

Kõige põhjalikum register Eesti inimeste kohta on rahvastikuregister (RR), mis koondab Eesti kodanike, Eestis elukoha registreerinud Euroopa Liidu kodanike ja Eestis elamisloa või elamisõiguse saanud välismaalaste peamisi isikuandmeid. RRi kantakse kõik rahvastikusündmused, sh sünnid, surmad ja registreeritud elukohavahetus. (Siseministeerium, 2022)

Kuid alaliste elanike kogumi määramiseks rahvastikuregistrist ainuüksi ei piisa, sest RRis esineb nii üle- kui ka alakaetust. Nii 2000. kui ka 2011. aasta rahvaloenduste puhul on toodud esile loenduse tulemuse erinevust rahvastikuregistrist. Rahvastikuregistri põhjal on elanike arv olnud 40 000 — 50 000 võrra suurem kui loendusega saadud või loenduse põhjal arvutatud rahvaarv. Peamiseks põhjuseks on toodud registreerimata väljarännet ehk seda, et rahvastikuregister sisaldab inimesi, kes on tegelikult Eestist lahkunud ning kes ei ole oma lahkumisest rahvastikuregistrit teavitanud. Seega ei saa rahvastikustatistikat teha põhinedes pelgalt rahvastikuregistrile, mis on levinud meetod näiteks Põhjamaades. (Maasing ja Tiit, 2016; Meres, 2017; Valner, 2012)

Selleks, et alaliste elanike kogum oleks võimalik kokku panna siiski registripõhiselt, ilma näost-näkku rahvaloendusi korraldamata, töötati välja residentsuse indeks (Maasing, 2015; Maasing, Tiit ja Vähi, 2017). Kuigi residentsuse indeksi põhjal avaldati rahvaarv esmakordselt 01.01.2016 kohta, on indeks välja arvutatud ka eelnenud aastate kohta (alates 2013. aastast). Residentsuse indeksi toimimise aluseks on laiendatud üldkogum ehk kogum inimestest, kes on olnud Eestiga seotud: kes on kuulunud RRi (nende elukoht seal võib olla nii Eestis kui välismaal või üldse puududa — nö passiivne osa rahvastikuregistrist), esinenud teistes elumärkidega seotud registrites või osalenud rahvaloendusel. (Maasing, Tiit ja Vähi, 2017; Meres, 2017)

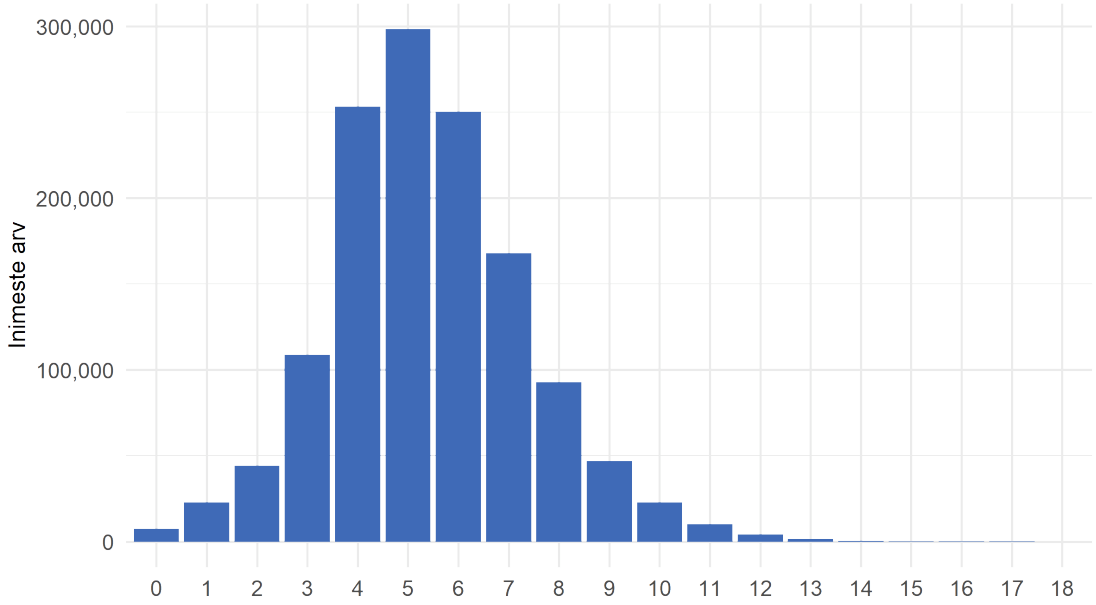
Üldkogumisse kuulub üle pooleteise miljoni inimese. Selleks, et indeksit arvutada on kasutusel 18 registrit, mille põhjal moodustub kokku 33 elumärki. Lisaks on kogumi moodustamiseks/elumärkide töötlemiseks kasutusel andmed veel kolmest registrist (EMSR — Sünniregister, SPR — Surma põhjuste register, RAKS — Riiklik rahvusvahelise kaitse andmise register). Elumärkide tuvastamiseks kasutatavad registrid ja neist moodustatavad elumärgid on toodud Tabelis 1. (Maasing, Tiit ja Vähi, 2017; Statistikaamet, 2022)

Tabel 1. Elumärkide moodustamiseks kasutatavad registrid.

Nr	Registri lühend	Registri nimi	Registri haldaja	Moodustuvad elumärgid
1	EHIS	Eesti Hariduse Infosüsteem	HTM	Õpilased; pedagoogid; huvihariduses osalejad
2	EMPIS	Töötuna ja tööotsijana arvel olevate isikute ning tööturuteenuste osutamise register	TK	Töötu või tööotsija
3	ESF	Euroopa Sotsiaalfondi tegevustes osalenud	SA	ESFi tegevustes osalenud
4	ET	E-toimik	JUM	Osalenud kohtuistungil, ülekuulamisel, toime pannud süüteo, isiklikult vastu võtnud dokumendi
5	ETR	Elamislubade ja töölubade register	PPA	Elamis- või tööluba
6	KIR	Vangide ja kriminaalhooldusaluste register	JUM	Viibinud kinnipidamisel või kriminaalhoolduses; külastanud kinnipeetavat
7	KIRST	Ravikindlustuse andmekogu	EHK	Hambaravi, raviarve; lapsepuhkus; töövõimetusleht; ravikindlustus
8	KMAIS	Isikut tõendavate dokumentide andmekogu	PPA	Vahetanud isikut tõendavat dokumenti
9	KVKR	Kaitseväekohustuslaste register	KRA	Kaitseväe- või asendusteenistuses
10	LR	Liiklusregister	TRAM	Teinud või vahetanud juhiluba; ostnud või müünud sõiduki
11	MKR	Maksukohustuslaste register	MTA	Maksnud makse
12	RETS	Retseptikeskus	EHK	Digiretsept
13	RR	Rahvastikuregister	SiM	Abielu; lahutus; elukoha muutus
14	SAP	Riigi personali- ja palgaarvestuse andmekogu	RTK	Riigiasutuses töötamine
15	SKAIS	Sotsiaalkaitse infosüsteem	SKA	Erihoolekanne; sotsiaaltoetus; pension; peretoetus; vanemahüvitis; puue ja/või töövõimetustõend
16	STAR	Sotsiaalteenuste ja -toetuste andmeregister	SKA	Saanud või taotlenud KOVilt sotsiaaltoetust
17	TETRIS	Töövõime hindamise ja töövõimetoetuse andmekogu	TK	Töövõime tõend
18	TÖR	Töötamise register	MTA	Eestis töötamine

Register või alamregister annab inimesele elumärgi siis, kui aasta jooksul on inimene teinud aktiivse sammu, mis on saanud kajastatud vastavas registris. Nii saab inimene elumärgi näiteks siis, kui ta on saanud sotsiaalhüvitisi, teeninud kaitseväes, olnud kohtus tunnistaja jne. Aasta jooksul omandab iga Eesti isikukoodiga inimene teatud hulga elumärke, aga võib ka olla, et elumärke ei kogune. (Maasing ja Tiit, 2016) Joonis 1

näitab, et keskmiselt kogus 01.01.2022 seisuga Eesti elanik 5 elumärki aastas.



Joonis 1. Kogutud elumärkide arv 2022. aasta residentidel.

Igale üldkogumisse kuuluvale inimesele j arvutatakse residentsuse indeksi väärtus $R(j,k)$, mis näitab tõenäosust, et isik j oli aastal k resident ehk alaline elanik. Residentuse indeks aastal k arvutatakse selle põhjal, mis oli residentsuse indeksi väärtus eelmisel aastal $k-1$ ja kui palju elumärke koguti eelmisel aastal $k-1$:

$$R(j,k) = dR(j,k-1) + gX(j,k-1) \tag{1}$$

kus kordajad d (stabiilsusmäär) ja g (elumärkide määr) omavad väärtusi $d = 0,8$ ja $g = 0,2$.

$$X(j,k) = \sum_{i=1}^{m} a_i E_i(j,k), a_i > 0 \tag{2}$$

on aastal k kogutud kaalutud elumärkide summa. Et kehtiks tingimus

$$0 \leq R(j,k) \leq 1, \tag{3}$$

on $R(j,k)$ kärbitud väärtuste 0 ja 1 vahel. (Maasing, Tiit ja Vähi, 2017)

Indeksi väärtused jäävad vahemikku [0, 1]. Mida suurem on indeksi väärtus, seda suurema tõenäosusega on isik Eesti resident. Kui $R(j,k) = 0$, siis j on kindel mitte-resident aastal k ning kui $R(j,k) = 1$, siis j on kindel resident aastal k. Kui indeksi väärtus jääb 0 ja 1 vahele, kasutatakse lävendit $c(0 \leq c \leq 1)$. Kui residentsuse indeksi väärtus ületab või on võrdne lävendiga arvestatakse need inimesed residentideks. Kui residentsuse indeksi väärtus on aga väiksem kui lävend, siis residendiks ei arvestata. Lävendiks on $c = 0,7$. (Maasing, Tiit ja Vähi, 2017)

Kindla residendi puhul säilib residendi staatus ka siis, kui ühe aasta jooksul ei ole elumärke kogunenud ($E_i(j) = 0, \forall i$), kuid mitte kauem. See tähendab, et kui kindel resident ei ole kahe aasta jooksul kogunud ühtegi elumärki, saab temast mitte-resident. Kindel mitte-resident saab residentsuse kuue aasta jooksul, kui ta on igal aastal saanud ühe elumärgi. Kui kindel mitte-resident saab ühel aastal vähemalt viis elumärki, saab ta residendiks juba järgmisel aastal. (Maasing, Tiit ja Vähi, 2017)

Kuna mõned elumärgid on rohkem residentsusega seostavad kui teised, on need kaalutud. Näiteks on erineva tähtsusega Eestis hooldekodus elamine ja juhtimisõiguse omandamine, mida võib saada ka inimene, kes on siia tulnud lühemaks perioodiks. Elumärkide kaalud arvutatakse eelmise aasta põhjal. Iga elumärgi kohta arvutatakse selle keskmine esinemine kindlate residentide ja kindlate mitte-residentide seas. Elumärgid, millel on kõrgem suhteline suurus, on tähtsamad. Kui lihtne elumärkide summa asendada kaalutud elumärkide summaga, on residendid ja mitte-residendid kergemini eristuvad. Selleks, et kaalud oleksid stabiilsemad, on kasutatud suhte logaritme. Indeks aitab paremini hinnata registreerimata sisserännet, kaasaarvatud inimeste naasmist, kes lahkusid registreerimata. (Maasing, Tiit ja Vähi, 2017)

Residentsuse indeksit arvutades võetakse arvesse ka rahvastikusündmused. Aastal $k-1$ sündinud või sisserände registreerinud inimeste indeksi väärtus on $R(k) = 1$. Sisserännanute puhul on oluline määrata ka elukoht Eestis. Registreeritult sisserännanud inimese puhul võetakse elukoht rahvastikuregistrist või eelmiste loenduse andmetest. Kui sisseränne on registreerimata ja inimesel ei ole elukohta, liigitub ta ootel sisserände alla. See tähendab, et vaadeldaval aastal ta residendiks ei saa (ta ei kvalifitseeru püsivaks residendiks), kuid kui tema residentsuse indeksi väärtus on ka järgneval aastal 1 ($R(k+$

$1) = 1$), siis arvestatakse ta residendiks, kelle elukoht on teadmata. Kui inimene aga ametlikult lahkub riigist, on tema indeksi väärtus $R(k) = 0$, kuid ta jääb endiselt alles laiendatud üldkogumisse ehk potentsiaalsete residentide hulka tulevikuks. Kui inimene sureb, siis tema potentsiaalsete residentide hulgast eemaldatakse. (Maasing ja Tiit, 2016; Maasing, Tiit ja Vähi, 2017)

1.3 Alaliste elanike määramise meetodid Euroopa riikides

Statistika kogumine oma inimeste, toodangu ja maa kohta pärineb tsivilisatsiooni algusaegadest. Esimene teadaolev rahvaloendus toimus Babüloonias aastal 3800 eKr. Rahvaloendused olid valitsejatele väga olulised infoallikad, mille järgi prognoositi, kui palju toitu ja tööjõudu on vaja, kui palju peaks makse koguma ja kui suur saaks olla sõjavägi. (Grajalez *et al.*, 2013)

Tänapäeval Euroopa riikidel ühtset viisi iga-aastase alaliste elanike kogumi kokkupanekuks ei ole ning riigid kasutavad erinevaid meetodeid. Suurem osa Euroopa Liitu või Euroopa Majanduspiirkonda kuuluvatest riikidest (16/32st) toodud Tabelis 2 hindavad rahvastiku suurust lähtudes rahvaloendusest. Rahvastikuregistri baasil paneb rahvaarvu kokku 11 riiki. Itaalia kasutab nii rahvaloendust kui ka rahvastikuregistrit. 4 riiki — Eesti, Bulgaaria, Läti ja Šveits panevad rahvaarvu kokku mitmete erinevate registrite baasil. (Eurostat, 2023)

Kuigi Euroopa Liidu rahva- ja eluruumide loendust puudutavate regulatsioonide järgi on eelistatud elanikkonna definitsioon alaline elanikkond (*usually resident population*), siis tulenevalt meetodite erinevusest on kasutusel ka teisi definitsioone. Üldjoontes on kasutusel 4 erinevat definitsiooni:

1. *de jure* elanikkond — põhineb isiku õiguslikul alusel riigis viibida; seega katab kõiki inimesi, kellel vastaval kuupäeval on kodakondsus, elamisluba või viisa
2. *de facto* elanikkond — kõik inimesed, kes on loendamise ajal kohapeal hoolimata sellest, kas neil on elukoht
3. registreeritud elanikkond — kõik inimesed, kes on olemas ühes või mitmes riikliku asutuse hallatavas registris referentskuupäeval

Tabel 2. Alaliste elanike määramise alus Euroopa riikides.

Riik	**Rahvastikuregister**	**Rahvaloendus**	**Muu**
Austria	X		
Belgia	X		
Bulgaaria			Rahvastikuregister, siseministeerium, riigi maksuamet
Eesti			Rahvastikuregister ja veel 17 Eesti haldusregistrit
Hispaania		X	
Horvaatia		X	
Iirimaa		X	
Itaalia	X	X	
Kreeka		X	
Küpros		X	
Läti			Matemaatiline meetod, Rahvastikuregister ja andmed muudest haldusallikatest
Leedu		X	
Luksemburg	X		
Madalmaad	X		
Malta		X	
Poola		X	
Portugal		X	
Prantsusmaa		X	
Rootsi	X		
Rumeenia		X	
Saksamaa		X	
Slovakkia		X	
Sloveenia	X		
Soome	X		
Taani	X		
Tšehhi		X	
Ungari		X	
Island	X		
Lichtenstein	X		
Norra	X		
Šveits			Piirkonna ja kohalike elanike registrid ja isikute föderaalne register
UK		X	

4. alaline elanikkond — inimesed, kes a) elasid enne referentskuupäeva riigis 12 kuud või b) saabusid 12 kuu jooksul enne refrerentsaega kavatsusega jääda sinna vähemalt 12 kuuks. See definitsioon on kasutusel Eestis ning see on ka

Euroopas kõige rohkem kasutatud definitsioon. (Eurostat, 2015) Samuti lähtub sellest definitsioonist uus Euroopa Komisjoni ettevalmistatav rahvastikustatistika määrus (Euroopa Komisjon, 2021).

Avaldamine toimub riikides enamasti iga-aastaselt, kuid on ka riike, kus rahvaarvu puudutavat statistikat avaldatakse juba praegu tihemini kui kord aastas. Üldiselt on need avaldamised vähem detailsemad kui aasta lõpu seisuga toimuvad avaldamised. (Eurostat, 2015) Väljaspool Euroopat on registripõhisele statistikale aluseks olevad universaalsete identifikaatoritega registrid haruldased. Paljudes riikides on nii tehnilised kui poliitilised takistused selliste süsteemide loomiseks (Lothian, Holmberg ja Seyb, 2019).

Põhjamaad. Nagu välja toodud, jagunevad riigid peamiselt rahvaloenduspõhise süsteemi ja registripõhise statistilise süsteemi vahel. Rahvaloenduspõhine süsteem on traditsiooniline vorm, kus küsitlejad koguvad rahvastiku ja eluruumide kohta andmeid iga 10 aasta tagant. Rahvaloenduste läbiviimine sellisel kujul on aga kallis ja keeruline. Samuti on loenduspõhine süsteem vähem informatiivsem. Kui administratiivsed süsteemid katavad kogu aega, siis rahvaloendused näitavad rahvastiku seisu ainult iga 10 aasta tagant. Nii näiteks ei saa tuvastada, millal täpsemalt mingi trend alguse sai ning milline on olukord praegusel hetkel. (A. Wallgren ja B. Wallgren, 2022)

Põhjamaid (Taani, Soome, Norra, Rootsi) võib pidada registripõhise statistilise süsteemi eestkõnelejaks. Neil riikidel on pikk traditsioon registrite kasutamisel ametlikus statistikas. Põhjamaade statistikaametid hakkasid registreid statistika tegemiseks kasutama 1960ndatel. Taani asendas klassikalise rahvaloenduse registripõhise loendusega juba 1981. aastal, Soome 1990. aastal ning Rootsi ja Norra 2011. aastal. (A. Wallgren ja B. Wallgren, 2022)

Taanis jäi 1981. aastal toimunud rahva- ja eluruumide loendus sellisel kujul ka viimaseks, sest rahvastiku ja eluruumide kohta hakati registripõhiselt statistikat avaldama iga-aastaselt. Lange (2014) viitab, et registripõhises statistilises süsteemis nagu Taani kaotas rahvaloendus esialgsel kujul oma mõtte. Taani väljaarendatud registripõhine metoodika on tunnustatud ka Euroopa Liidu tasandil 1995. aastast (Euroopa Komisjon, 1995).

Registripõhise statistilise süsteemi loomist on neis riikides toetanud õiguslik alus

ja avalikkuse toetus. Õiguslik alus määrab, milliseid andmeid toodetakse, kas inimesed peavad oma kolimisest, riiki saabumisest või riigist lahkumisest registrit teavitama. Samuti määrab õiguslik alus selle, kas statistikaamet saab registriandmeid kasutada. Näiteks Taani süsteemi tekke võtmeelement oli see, et seadusandjad suhtusid algusest peale väga positiivselt registripõhisesse statistikasse (Euroopa Komisjon, 1995).

Põhjamaade statistiliste registrite süsteem koosneb mitmetest registritest, mida saab üksteisega siduda. Kuid põhiliseks aluseks on rahvastikuregister, mida uuendatakse igapäevaselt. Põhjamaades on pikk traditsioon anda riiklikule registripidajale teada, kui kolitakse uuele aadressile. See on kohustuslik ja paljud sotsiaaltoetused on seotud selle omavalitsusega, kus elatakse. Seega uuendatakse registris infot, kui eluaset vahetatakse. Niisiis loob administratiivne süsteem igapäevaselt statistikat, mis on võrreldav lühemas formaadis rahvaloendusega. Kuna kõik asutused kasutavad rahvastikuregistrit, siis saavad ka kõik asutused, pangad ja kindlustusfirmad automaatselt elukohavahetusest teavitatud. Seega on elanikel stiimul, et hoida oma andmeid uuendatuna, et saada informatsiooni, sotsiaaltoetusi ja teisi sotsiaalseid teenuseid ja garantiisid. Administratiivset registrit uuendatakse ka näiteks ametkondadelt, tervishoiutöötajatelt ja üksikisikutelt tuleva info põhjal. (A. Wallgren ja B. Wallgren, 2022)

Rahvastikuregistrile toetumises on ka negatiivseid külgi. Näiteks on neis riikides elanikkonna definitsioon reeglina teistsugune, kui Euroopa Liidus soovituslik alaline elanikkond. Põhjamaades on rahvastikus ainult inimesed, kes on registris registreeritud ehk lähtutakse registreeritud elanikkonna definitsioonist. See teeb statistika tootmise lihtsaks, kuid inimesed, kes tulevad neisse maadesse tööle või õppima, jäävad alaesindatud ehk alakaetuse probleem on süvenemas. Samuti on paljud välismaalased registreeritud kui ajutised elanikud ning välja jäävad ka illegaalsed immigrandid. Ka neis riikides soovitakse selle tõttu kaasata rohkem näiteks ülikooli- ja maksuregistreid. Samuti võivad väljarändajate riigist väljaregistreerimised mitte saada kajastatud õigeaegselt ja täpselt. (A. Wallgren ja B. Wallgren, 2022)

Kokkuvõttes on Eesti alaliste elanike kokkupaneku metoodika üks erilisemaid, kuid ka andmemahukamaid. Uue määruse väljatöötamine annab meile võimaluse muuta oma metoodikaid innovatiivsemaks. Samas on oluline jälgida, et uued metoodikad ei süvendaks ala- ega ülekaetust.

2 Metoodika

See peatükk annab ülevaate sellest, millist metoodikat kasutati töö eesmärgi täitmiseks ja uurimisküsimustele vastamiseks. Täpsemalt, kuidas vähendati tunnuste hulka ning milliseid masinõppe mudeleid testiti, et hinnata nende sobivust residentsuse määramiseks.

Töös on kasutatud Statistikaameti poolt selle teadustöö tarbeks kättesaadavaks tehtud andmeid (Statistikaamet, 2023). Eksperimentide läbiviimiseks on kasutatud programmeerimiskeele R versioone 4.1.3 ja 4.0.5 ning kasutajaliidest RStudio. Kasutatud on lisapakette nagu *tidyverse* (versioon 1.3.2, 1.3.1) (Wickham *et al.*, 2019), *caret* (6.0-93) (Kuhn, 2008), *pROC* (1.18.0) (Robin *et al.*, 2011), *e1071* (1.7-11) (Meyer *et al.*, 2023), *rpart* (4.1.16) (Therneau ja Atkinson, 2022), *randomForest* (4.7-1.1) (Liaw ja Wiener, 2002) ja *xgboost* (1.7.5.1) (Chen *et al.*, 2023).

Töös on kasutatud OpenAI arendatud keelemudelit ChatGPT (GPT-3.5), et saada metoodilisi juhiseid ja nõu lähtekoodi kirjutamiseks (OpenAI, 2023). ChatGPT poolt saadud vastuseid on kontrollitud teiste allikate põhjal.

2.1 Andmed

Eksperimentide läbiviimiseks kasutati 2022. aastal residentsuse arvutamiseks koostatud andmeid. Need andmed on Statistikaametis kokku pandud residentsuse indeksi arvutamiseks 01.01.2022 seisuga. Nimelt on Eesti elanike kogumi kokkupanemisel tehtud andmestik, kus on nii residentsuse määramiseks olulised muutujad kui ka tulemused. Seal on residentsust puudutav info nii praegu kui ka varasemalt Eestiga seotud isikute kohta ehk kogu üldkogumi kohta, mis 2022. aastal koosnes 1 826 218 kirjest. Kõige suurem hulk inimesi on tulnud sinna erinevate aastate (2012—2022) RRi väljavõtetest — 93%. Samuti on inimesi tulnud üldkogumisse teiste elumärkidega seotud registritest — 6%. 2011. ja 2000. aasta rahvaloendusest ja sündide registritest on üldkogumisse tulnud lisaks 1% kirjetest.

Esialgne andmestik koosnes kuuekümne neljast tunnusest, millest kolmkümmend kolm olid elumärkide tunnused — kas inimene on olnud vastavas spetsiifilises registris vaatlusaastal või mitte. Kuus tunnust tulenesid Rahvastikuregistrist — kas inimene

on olnud vaatlusaastal või sellele eelnenud kahel aastal RRi väljavõttes ning milline on tema elukoht seal (Eesti, välismaa või puudub). Samuti on andmestikus üheksa üldist kirjeldavat tunnust. Kuusteist tunnust on selle ja eelneva kolme aasta residentsuse arvutamiseks vajalikud tunnused või selle tulemusena saadud tunnused, nagu kogutud elumärkide arv, residentsuse indeksi väärtus, residentsuse määramise põhjendus jne.

Ennustatav tunnus — residentsus — on 01.01.2022 seisuga residentsus (0: mitteresident, 1: resident). Tegemist on residentsuse indeksi metoodika käigus arvutatud väärtusega, seega ei saa öelda, et tegemist oleks kindla tõega, küll aga on varasem väärtus kasulik mudeli treenimiseks ja täpsuse hindamiseks. Kirjetest 1 331 796 juhul on residentsuse väärtus "1"(rahvastik seisuga 01.01.2022) ja 494 422 väärtus "0".

Andmete eeltöötlemise käigus eemaldatakse 2022. aasta tulemusega seotud muutujad, v.a residentsuse väärtus, millest saab mudelites sõltuv muutuja. Lisaks eemaldatakse üldkogumist kõik surnud inimesed, eeldades, et tulevikus sarnast mudelit kasutades treenitakse mudeleid samuti andmestikul, kuhu on juba lisatud vaatlusaasta sünnid ja eemaldatud surnud. Sündide ja surmade andmed on faktilised ja tulevad registritest, seega ei ole vajalik neid sündmusi mudeliga ennustada. Samuti eemaldatakse näiteks üksteist dubleerivad tunnused ehk tunnus, mille korrelatsioon mõne teise tunnusega on 1. Kuus kategoorilist tunnust muudetakse *one-hot encoding* abil binaarseteks tunnusteks. Tühjad väärtused muudetakse nulliks. Andmestik segatakse ära ja kõik muutujad peale sõltuva muutuja normaliseeritakse kasutades R-i funktsiooni *scale*, mis normaliseerib andmed nii, et tunnuste keskmine väärtus oleks 0 ja standardhälve 1.

Andmestikku jääb lõpuks 74 muutujat ja 1 619 722 kirjet. Andmestik jaotatakse treening- ja testandmestikuks, kus 70% andmetest valiti juhuslikult treeningandmestikku (1 133 660 vaatlust) ja 30% testandmestikku (486 062 vaatlust). Mudelite treenimiseks jaotati treeningandmed omakorda kaheks: treeningandmed (80% — 906 923 kirjet) ja valideerimisandmed (20% — 226 737 kirjet).

2.2 Tunnuste selekteerimine

Töö eesmärgi täitmiseks kasutatakse eeltöötlusetapina dimensioonide vähendamist (ingl k *dimensionality reduction*), mis aitab kõrvaldada ebaolulised andmed, müra ja

üleliigsed tunnused. See lihtsustab andmete analüüsi, töötlemist ja visualiseerimist. Siin töös kasutatakse selle alammeetodit tunnuste selekteerimist (ingl k *feature selection*). Tunnuste selekteerimine tähendab vaatamist, kas kõik olemasolevad tunnused on tegelikult kasulikud, st on väljundmuutujatega korrelatsioonis. Selle abil saab olenevalt määratletud kriteeriumidest valida asjakohaste tunnuste alamloendi, mis kirjeldaks võimalikult adekvaatselt kogu andmeid. Tunnuste selekteerimine aitab vähendada andmestiku suuremõõtmelisust ja klassifitseerival algoritmil keskenduda olulisematele tunnustele, ignoreerides eksitavaid tunnuseid. See aitab ka paremini aru saada tunnuste ja ennustatava muutuja suhtest vähendades arvutusvajadust. (Marsland, 2014)

Tunnuste valimise meetodina kasutati peakomponentide analüüsi (PCA, ingl k *principal component analysis*). See on levinud ja lihtsasti kasutatav meetod, mis sobib kasutamiseks suurel andmehulgal. PCA idee seisneb selles, et koordinaattelgede komplektide leidmisel saab selgeks, et mõned mõõtmed pole vajalikud. Peakomponentide analüüs leiab tunnuste lineaarsed kombinatsioonid ehk komponendid, mis säilitavad tunnustest kõige olulisema informatsiooni kirjeldades kõige enam andmete variatiivsust. Komponendid, mis kirjeldavad vähe variatiivsust saab välja jätta. (Marsland, 2014) Peakomponentide analüüsi rakendati treeningandmetel, et tuvastada olulisimad tunnused, mida kasutada residentsuse arvutamiseks.

Kõige tähenduslikumad tunnused määrati kahe esimese peakomponendi põhjal. Valiti tunnused, mille laadungite absoluutväärtus oli suurem kui 0,1. Ehk valituks osutusid tunnused, mis olid kahes esimeses peakomponendis kõige olulisemad.

2.3 Masinõppe mudelid

Peale olulisimate tunnuste kindlakstegemist rakendati andmetel masinõppe meetodeid. Masinõpe on meetodite kogum, mis suudab automaatselt tuvastada andmetest mustreid ning kasutada leitud mustreid, et ennustada tuleviku andmeid või teostada ebakindlas kontekstis mingit muud otsustusprotsessi (Murphy, 2012). Masinõpet jaotatakse juhendatud ja juhendamata õppeks. Ennustava ehk juhendatud õppe eesmärk on õppida sisendi x ja väljundi y vaheline suhe, kui on olemas kogum märgendatud sisendi-väljundi paare. Kui y on kategooriline, on üldiselt tegemist klassifitseerimisülesandega ning kui y on arvuline, on tegemist regressiooniga. (Murphy, 2012)

Selles töös kasutatav andmestik põhineb juba arvutatud eelmise aasta residentsuse indeksile, seega on olemas märgendatud andmed (resident, mitteresident) ning saab kasutada juhendatud õppe meetodeid. Kuna on kaks klassi, on tegemist binaarse klassifikatsiooniga. Samas on residentsuse indeks arvutuslik meetod, mis ei pruugi kajastada täielikku tõde. Taolistel ebamäärastel juhtudel on kasulik kasutada meetodeid, mis väljastavad tõenäosuse ehk tõenäosuslikke klassifitseerimismeetodeid, mis annavad pigem prognoositava märgendi tõenäosuse hinnangu kui lõpliku klassifikatsiooni. Sel juhul saab mudel määrata igale võimalikule klassisildile tõenäosusskoori ja prognoositava klassisildi määramiseks saab kasutada lävendit. (Murphy, 2012)

2.3.1 Logistiline regressioon

Logistiline regressioon on statistiline mudel, mille parameetrite arv on fikseeritud. Seega on see mudel vähem paindlik kui mõned teised, samas on see efektiivne suuredimensioonilistel andmetel kasutamiseks. Logistilise regressiooni kasutamine on otstarbekas, kui andmestiku sõltuv muutuja on binaarne. Logistiline regressioon kasutab sigmoidfunktsiooni, et väljund jääks 0 ja 1 vahele. Otsuse tegemiseks saab määrata tõenäosuse künnise ehk otsustusreegli. (Murphy, 2012) R-is kasutati logistilise regressiooni mudeli treenimiseks baas R-is olevat funktsiooni *glm()*.

2.3.2 Naiiv-Bayes

Naiiv-Bayesi klassifikaator on tõenäosuslik meetod, mis kasutab Bayesi teoreemi eeldusega, et ühe tunnuse olemasolu selles klassis ei ole seotud teise tunnuse olemasoluga samas klassis. Naiiv-Bayes eeldab, et kõik andmestiku tunnused on võrdselt olulised ja sõltumatud. Kategooriate ühiseid tõenäosusi kasutatakse antud kategooriate tõenäosuste ennustamiseks. Selle sõltumatuse eelduse tõttu saab iga liikme parameetreid eraldi uurida, seega kiirendab see arvutustoiminguid. Bayesi võrk koosneb struktuurmudelist ja tingimuslike tõenäosuste komplektist. (Lantz, 2019) R-is kasutati Naiiv-Bayesi mudeli treenimiseks paketi *e1071* (1.7-11) funktsiooni *naiveBayes()* (Meyer *et al.*, 2023).

2.3.3 Otsustuspuu

Otsustuspuu on levinud juhendatud õppemeetod, mida kasutatakse nii regressiooni- kui ka klassifitseerimisprobleemide lahendamiseks. Selle eesmärk on tulemuse prognoosimine lihtsate otsustusreeglite abil. Otsustuspuude meetodit on lihtne tõlgendada ning kasutada, kuid liiga keeruliste puude ehitamine, mis põhjustab ülepaigutamist, on tihti nõrk külg. (Lantz, 2019) R-is kasutati otsustuspuu mudeli treenimiseks paketi *rpart* (4.1.16) funktsiooni *rpart()* (Therneau ja Atkinson, 2022).

2.3.4 Otsustusmets

Otsustusmetsa klassifikaator kuulub koosmõju (ingl k *ensemble*) klassifikaatorite rühma. Paremate ennustustulemuste saamiseks kasutab see paralleelseid otsustuspuu mudeleid. See loob palju puid ja igale puule rakendatakse *bootstrap* tehnikat. Klassifikatsioonis antakse protseduuri sisend igale metsas olevale puule ja seejärel hääletab iga puu eraldi selle klassi poolt. Lõpuks valib otsustusmets välja klassi, mis on saanud kõige rohkem hääli. (Hearty, 2016) R-is kasutati otsustusmetsa mudeli treenimiseks paketi *randomForest* (4.7-1.1) funktsiooni *randomForest()* (Liaw ja Wiener, 2002).

2.3.5 XGBoost

Koosmõju klassifikaatorid saavad põhineda ka *boosting* mudelile, nagu XGBoost. *Boosting* meetodit iseloomustab järjest mitmete mudelite kasutamine, et korduvalt tõsta ("*boost*") ja parandada koosmõju toimimist. 2015. aastal loodud *Extreme Gradient Boosting* (XGBoost) algoritm on näidanud häid tulemusi erinevate andmeteaduslike ülesannete lahendamisel. Iga kordusega püüab XGBoost parandada mudelit vähendades koosmõjude jääki. (Hearty, 2016) R-is kasutati XGBoost mudeli treenimiseks paketi *xgboost* (1.7.5.1) funktsiooni *xgboost()* (Chen *et al.*, 2023).

2.3.6 Mudelite hindamine

Mudelite headuse hindamiseks mõõdetakse klassifitseerimistäpsus testandmestikul, mis on 30% algsest andmestikust. See võimaldab hinnata, milline oleks mudeli täpsus uutel

andmetel. Kuna testandmeid saab kasutada alles viimase sammuna, on treeningandmestik jaotatud 80% treeningandmeteks ja 20% valideerimisandmeteks, et hinnata mudeli täpsust eelkõige erinevatel tõenäosuskünnistel.

Tõenäosuskünnis hinnati valideerimisandmete põhjal kasutades Youden'i indeksit. Ehk siis leiti punkt, kus tegelik positiivsus ja tegelik negatiivsus on parimas suhtes ehk Youden'i indeks on maksimaalne. R-is kasutati selle punkti määramiseks paketi *pROC* (1.18.0) funktsiooni *coords()* (Robin *et al.*, 2011).

Erinevate mudelite võrdlemiseks ja headuse hindamiseks kasutati nelja mõõdikut ning lisaks võrreldi segadusmaatriksite tulemusi. Neli mõõdikut, mida kasutati on ühed kõige tihemini kasutatavad hindamismõõdikud: korrektsus (ingl k *accuracy*), täpsus (ingl k *precision*), saagis (ingl k *recall*) ja F1-skoor. Nii need neli näitajat kui ka segadusmaatriks tuginevad neljale väärtusele: tegelik positiivsus, tegelik negatiivsus, valepositiivsus ja valenegatiivsus.

Tegelik positiivsus tähendab, et kui tegelik klass (väärtus) on 1, siis on ka ennustatav klass 1. Tegelik negatiivsus tähendab, et kui tegelik klass on 0, siis on ka ennustatav klass 0. Valenegatiivsus ja valepositiivsus esinevad siis, kui ennustatav ja tegelik klass erinevad: valenegatiivsuse puhul on tegelik klass 1, aga ennustatav 0 ja valepositiivsel juhul tegelik klass 0, aga ennustatav klass 1. Korrektsus on õigesti ennustatud vaatluste koguarvu suhe vaatluste koguarvuga. (Lantz, 2019) Segadusmaatriksi ülesehitus selle töö kontekstis on toodud Tabelis 3.

Tabel 3. Segadusmaatriksi ülesehitus.

	Tegelik mitteresident	**Tegelik resident**
Ennustatud mitteresident	Tegelik negatiivne (TN)	Valenegatiivne (FN)
Ennustatud resident	Valepositiivne (FP)	Tegelik positiivne (TP)

Korrektsust, täpsust, saagist ja F1-skoori saab arvutada järgmiste valemite järgi:

$$\text{Korrektsus} = \frac{TP + TN}{TN + TP + FP + FN} \tag{4}$$

$$\text{Täpsus} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Saagis} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1-skoor} = \frac{2 \times Täpsus \times Saagis}{Täpsus + Saagis} \tag{7}$$

3 Tulemused

Järgnev peatükk annab ülevaate töö tulemustest. Töö uurimisprobleemi lahendamiseks kasutati mitmeid statistilisi ja masinõppe meetodeid. Olulisemate tunnuste valimiseks kasutati peakomponentide analüüsi ning saadud andmestikel testiti erinevaid masinõppe mudeleid, et hinnata nende sobivust residentide määramiseks.

3.1 Valitud tunnused

Selle magistritöö üks peamisi eesmärke oli uurida, millised muutujad on residentsuse määramisel kõige olulisemad, selleks, et oleks võimalik vähendada sisendandmete hulka ning seeläbi võimaldada kiiremat ja efektiivsemat rahvastikustatistika kokkupanemist. Töö üheks uurimisküsimuseks oli: „Millised tunnused praegu olemasolevatest andmetest on kõige vajalikumad, et määrata residentsust?“ Olulisemate tunnuste valimiseks kasutati peakomponentide analüüsi meetodit. 73-tunnuse põhjal tehtud peakomponentide analüüsis kirjeldasid kaks esimest peakomponenti 22% andmetes esinevast variatiivsusest. Järgmiseks leiti konkreetsed tunnused, mis olid kahes esimeses peakomponendis kõige olulisemad ehk mille laadungi absoluutväärtus oli suurem kui künnis 0,1. Peakomponentide analüüsi läbiviimiseks kasutati R-i funktsiooni *prcomp*.

Peakomponentide analüüs põhjal selgus, et andmestikus saab kõige olulisemateks pidada kolmekümmend tunnust. Valitud tunnused on toodud Tabelis 4. Üheksa neist on seotud erinevate valdkondlike elumärkidega: õppimine (EHIS), Eestis töötamine (TÖR), sotsiaaltoetuse saamine, pensioni saamine, peretoetuste saamine, vanemahüvitise saamine, raviarve omamine, digiretsepti omamine ja maksude maksmine vaadeldaval perioodil. Üheksa tunnust on aga seotud rahvastikuregistriga, nimelt omab olulist kaalu esinemine rahvastikuregistris nii vaatlusaastal kui kahel eelneval aastal ning Eesti või välismaa elukoha omamine rahvastikuregistris nii vaadeldaval kui kahel eelneval aastal. See tulemus näitab, et olulisemad tunnused on peamiselt seotud töötamisega, õppimisega, sotsiaaltoetustega ja meditsiiniteenustega ning samuti on oluline rahvastikuregister.

Üheksa tunnust on seotud residentsusega eelnevatel aastatel — kogutud elumärkide arv eelmisel aastal, residentsus eelmisel ja üle-eelmisel aastal, residentsuse indeksi väärtus enne ümardamist eelmisel ja üle-eelmisel aastal; fakt, et eelmise- ja üle-eelmise

Tabel 4. PCA selekteeritud tunnused algandmestikust.

Nr	Tunnus	Nr	Tunnus
1	Eesti isikukoodi olemasolu (0/1)	16	Indeksi väärtus enne ümardamist 2021
2	RRi elukoht Eestis 2022 (0/1)	17	Residentsus aastal 2020 (0/1)
3	RRi elukoht välismaal 2022 (0/1)	18	Residentsus 2020 tulenes indeksist (0/1)
4	RRi elukoht Eestis 2021 (0/1)	19	Residentsus 2020 kuna ema oli resident (0/1)
5	RRi elukoht välismaal 2021 (0/1)	20	Indeksi väärtus enne ümardamist 2020
6	RRi elukoht Eestis 2020 (0/1)	21	Sünniaasta
7	RRi elukoht välismaal 2020 (0/1)	22	EHISe elumärk (0/1)
8	RRis esinemine 2022 (0/1)	23	TÖRi elumärk (0/1)
9	RRis esinemine 2021 (0/1)	24	Sotsiaaltoetuse elumärk (0/1)
10	RRis esinemine 2020 (0/1)	25	Pensioni elumärk (0/1)
11	Ema isikukoodi olemasolu (0/1)	26	Peretoetuse elumärk (0/1)
12	Kogutud elumärkide arv 2021 (0/1)	27	Vanemahüvitise elumärk (0/1)
13	Residentsus aastal 2021 (0/1)	28	Raviarve elumärk (0/1)
14	Residentsus 2021 tulenes indeksist (0/1)	29	Digiretsepti elumärk (0/1)
15	Residentsus 2021 kuna ema oli resident (0/1)	30	Maksude elumärk (0/1)

aasta residentsus oli saadud arvutades (mitte tulenevalt registreeritud sündmusest nagu väljaränne, sisseränne, surm, vangistus, sündinutel ema staatus). See on loogiline tulemus arvestades tõsiasja, et suurem osa Eesti elanikest on siinsed elanikud püsivalt, näiteks 2022. aasta residentidest olid residendid ka 2021. aastal 97,5%. Samuti on olulisteks tunnusteks sünniaasta, Eesti isikukoodi olemasolu ja ema isikukoodi olemasolu, mis tuleb peamiselt RRist ja/või sündide registritest.

Andmehõive seisukohast hõlmab saadud tunnuste koguarv seega seitset erinevat registrit — RR, EHIS, TÖR, SKAIS (sotsiaaltoetused, pensionid, peretoetused, vanemahüvitised), KIRST (raviarve), RETS (digiretseptid) ja MKR (maksude maksmine). Lisaks on baaskogumile juba lisatud vaatlusperioodil toimunud sünnid ja eemaldatud surmad. Siiski on seitsme registri põhjal residentide määramine ressursisäästlikum kui 18 registri põhjal.

3.2 Mudelite võrdlus

Järgnevalt filtreeriti andmestikust välja kolmkümmend eelnevalt leitud muutujat ning lisaks sõltuv muutuja — residentsus 2022. aastal. Saadud andmestiku peal rakendati viit erinevat masinõppe mudelit: logistiline regressioon, Naiiv-Bayes,

otsustuspuu, otsustusmets ja XGBoost. Mudelite puhul jäid aluseks vastavate R-i funktsioonide vaikeparameetrid, v.a XGBoost, kus iteratsioonide arvuks määrati 1000, mis parandas mudeli täpsust, kuid ei suurendanud oluliselt treenimiseks kuluvat aega. Valideerimistulemuste põhjal määrati tõenäosuskünnis Youden'i indeksi põhjal. Tabelis 5 on toodud igale mudelile rakendatud tõenäosuskünnised.

Tabel 5. Mudelite tõenäosuskünniste väärtused.

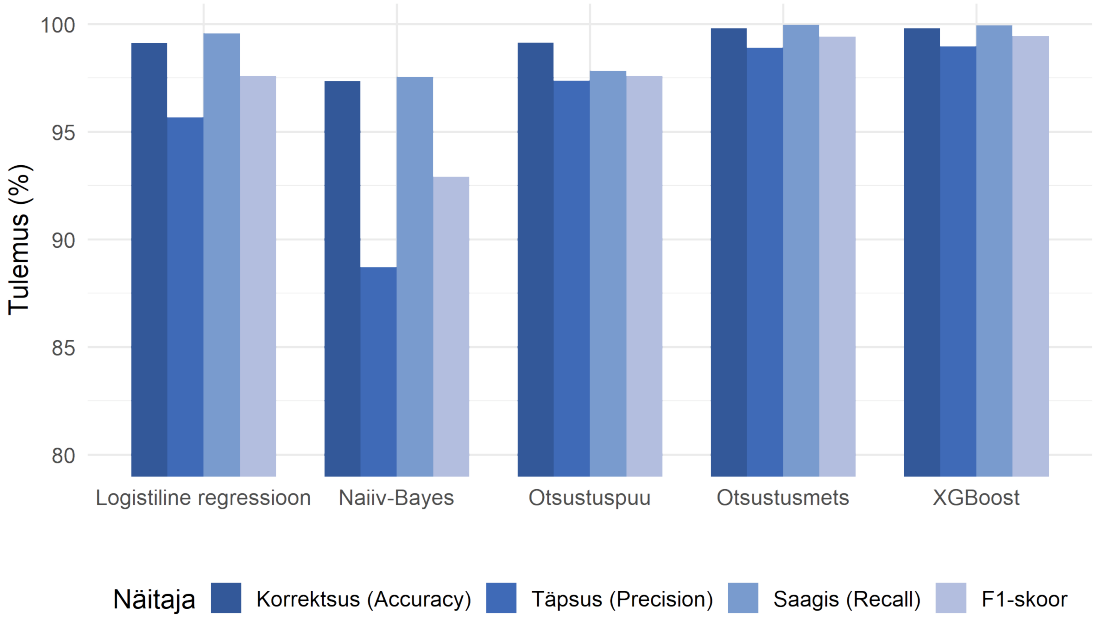
Algoritmid	**Tõenäosuskünnis**
Logistiline regressioon	0,8
Naiiv-Bayes	1
Otsustuspuu	0,5
Random Forest	0,6
XGBoost	0,9

Mudelite võrdlemiseks kasutati nelja kriteeriumi: korrektsus, täpsus, saagis ja F1-skoor. Erinevate algoritmide tulemused testandmestikul on toodud Tabelis 6 ja graafiliselt Joonisel 2.

Tabel 6. Erinevate masinõppe mudelite tulemus testandmetel.

Algoritmid	**Korrektsus (Accuracy)**	**Täpsus (Precision)**	**Saagis (Recall)**	**F1-skoor**
Logistiline regressioon	99,12%	95,68%	99,57%	97,59%
Naiiv-Bayes	97,35%	88,71%	97,55%	92,92%
Otsustuspuu	99,14%	97,37%	97,82%	97,59%
Otsustusmets	**99,80%**	98,91%	**99,96%**	99,43%
XGBoost	**99,80%**	**98,96%**	99,94%	**99,45%**

Nagu tabelist ja jooniselt nähtub, määrasid residentsust kõige paremini otsustusmets ja XGBoost, mis on koosmõjul põhinevad algoritmid. Samas on korrektsuse ja saagise osas üsna võrdväärne ka logistiline regressioon, mis on mudelina lihtsamini seletatav ja kiirem kui näiteks otsustusmets. Kõigi mudelite puhul on kõige madalamaks näitajaks täpsus ehk tegelike positiivsete tulemuste osakaal kõigist positiivsetest ennustustest. Kõige kehvemat tulemust näitas Naiiv-Bayes, mida saab põhjendada sellega, et Naiiv-Bayes eeldab tunnuste omavahelist sõltumatust. Kui võtta arvesse mudeli treenimiseks kulunud aega, siis tuleks eelistada XGBoost mudelit, mille treenimine 1000 iteratsiooniga võttis aega ligi neli minutit, otsustusmetsa algoritmi treenimine võttis aega aga üle poole tunni.



Joonis 2. Algoritmide tulemusnäitajad.

Mudelite korrektsuse võrdlus vähendatud (30 tunnust) ja algses andmestikus (73 tunnust) on toodud Tabelis 7. On näha, et vähendatud andmestiku korrektsus on lähedal algandmestiku korrektsusele. Ka algandmestikul toimib kõige paremini otsustusmetsa algoritm saades korrektsuseks 99,93% ning korrektsus vähendatud andmestikul jääb sellele alla vaid 0,13 protsendipunktiga.

Tabel 7. Mudelite korrektsus (*accuracy*) algandmestikus (73 tunnust) ja vähendatud andmestikus (30 tunnust).

Algoritmid	**Algandmestik**	**Vähendatud andmestik**
Logistiline regressioon	99,41%	99,12%
Naiiv-Bayes	97,19%	97,35%
Otsustuspuu	98,72%	99,14%
Random Forest	**99,93%**	**99,80%**
XGBoost	99,88%	**99,80%**

Lisaks eelnevatele kriteeriumitele analüüsiti tulemusi ka juhtumipõhiselt. Selleks kasutati segadusmaatriksit (ingl k *confusion matrix)*. See on oluline, sest põhiline

meetod, mille järgi mudelite täpsust parandada, on analüüsida valesti klassifitseeritud juhtumeid. Rahvastiku määramisel on see eriti oluline, et ei esineks ei valepositiivsust (inimene ei ole kriteeriumite järgi resident, kuid mudel ennustab residendiks) ega ka valenegatiivsust (inimene on kriteeriumite järgi resident, aga ennustatud mitte-residendiks). Segadusmaatriksi tulemused testandmetel on toodud Tabelis 8.

Tabel 8. Algoritmide tulemus testandmetel, inimeste arv.

Algoritm	**TP**	**FP**	**TN**	**FN**
Logistiline regressioon	395 663	368	86 138	3893
Naiiv-Bayes	388 814	2116	84 390	10 742
Otsustuspuu	397 273	1888	84 618	2283
Otsustusmets	398 600	**36**	**86 470**	956
XGBoost	**398 647**	54	86 452	**909**

Nagu Tabelist 8 selgub, on ka valepositiivsete ja valenegatiivsete hulk otsustusmetsa ja XGBoosti algoritmide puhul kõige väiksem. Kõige vähem valepositiivseid annab otsustusmetsa algoritm, samas kõige vähem valenegatiivseid annab XGBoost algoritm. XGBoosti ja otsustusmetsa puhul olid valepositiivsed juhud kattuvad 13 kirje puhul. Peamiseks ühendavaks teguriks valepositiivsete juhtude puhul saab tuua sündimise 2021. või 2020. aastal. Otsustusmetsa puhul on valepositiivsetest juhtudest 36% sündinud neil aastatel, XGBoosti puhul on varieeruvus suurem, kuid kõige suurem osakaal inimestest on sündinud 2021 — 11%. Siiski on valepositiivseid juhte kokkuvõttes üsna vähe.

Valenegatiivseid juhtumeid ehk juhtumeid, kus mudel on ennustanud mitteresidendiks, kuid residentsuse indeksi järgi oli inimene resident, on rohkem. Kattuvaid isikuid on otsustusmetsa ja XGBoosti valenegatiivsete hulgas 848 ehk 93% XGBoost valenegatiivsetest ja 89% otsustusmetsa valenegatiivsetest. Nende puhul on ühendavaks teguriks Eesti isikukoodi puudumine (93%), samuti KMAISi elumärk, mida vähendatud andmestikus ei olnud (42%). Samuti olid neist eelneval ehk 2021. aastal mitteresidendid 52% ning 2020. aastal 44% ehk tegemist ongi olnud pigem mittestabiilsete juhtumitega. Mudelist puudunud elumärkidest võib kõige määravam olla KMAIS, mida valenegatiivsetel juhtudel oli päris paljudel. Teiste elumärkide kohta seda pigem öelda ei saa, sest elumärgi olemasolu osakaalud jäävad valdavalt 10% juurde.

4 Arutelu

Kuigi masinõppe mudelite tulemused näitasid, et on võimalik üsna täpselt elanikkonda ennustada ka oluliselt väiksema arvu elumärkidega, on siiski oluline arvesse võtta, et elumärkide olulisus võib olla aastati/perioodi erinev. Eriliste olukordade puhul võivad oluliseks osutuda uued elumärgid. Näiteks suurema hulga põgenike riiki saabumine muudab tõenäoliselt olulisemaks KMAISi ja/või ETRi elumärgi (isikut tõendava dokumendi saamine, elamis- või tööloa saamine). Seetõttu on oluline analüüsida elumärkide olulisust ka edaspidi, eriti kui toimuvad ühiskondlikud muutused.

Tulevikus rahvastikustatistika kokkupanemisel tähendab see, et olulisemad tunnused tuleks määrata/üle kontrollida aasta alguses 1. jaanuari seisuga rahvaarvu moodustades ning siis saaks aasta keskel avaldamiseks kasutada neid tunnuseid, arvestades elumärgi olemasolu aasta keskpaiga seisuga. Ehk siis oleks tulevikus oluline võtta aasta keskel arvutatavale rahvastikule aluseks need tunnused, mis olid viimati arvutatud täieliku indeksi puhul olulisemateks, näiteks määrates residente 1. juuli seisuga 2022 tuleks aluseks võtta olulisemad tunnused 1. jaanuar 2022 põhjal. Niisiis võiks aasta keskpaigas toimuvaks avaldamiseks arvesse võtta poolt aastat senise aasta asemel. Kuid selle muutuse mõjusid tuleks eraldi hinnata — kas olulisemate elumärkide puhul erineb pooleaastane ja aastane vaatlusperiood, sest elumärgi saamise aeg väheneb. Samas on peakomponentide analüüsiga leitud olulisemad elumärgid üsna stabiilsed, näiteks töötamine, õppimine või toetuste saamine ning suur osa elanikkonnast ongi püsielanikud.

Analüüsiga leitud olulisemad tunnused haakuvad ka teistes riikides rahvastikuregistrile lisaks kasutatavad registritega. Ka A. ja B. Wallgren (2022) toovad rahvastikuregistri täiendustena välja töötamise, õppimise, maksude, toetuste ja meditsiiniteenustega seotud registrid. Samuti toovad nad välja registripõhise statistika toetamiseks ka küsitlusandmete kasutamist, näiteks tööjõu-uuring ja muud valikuuringud. Ka nende võimalikku kasutamist tulevikus residentsuse indeksi täiustamiseks ja kontrollimiseks saab kaaluda.

Kokkuvõte

See magistritöö keskendus sellele, kuidas muuta efektiivsemaks rahvastikustatistika kokkupanemist. Kasvav vajadus ajakohasema ja samas kvaliteetse statistika vastu muutub üha aktuaalsemaks. Hiljemalt 2026. aastal tuleb Euroopa Komisjoni määrusest tulenevalt Eestil rahvaarvu avaldada eeldatavasti kaks korda aastas. Praegune alaliste elanike kokkupaneku metoodika on kvaliteetne, kuid tulevikus võib vaja olla dünaamilisemat lahendust, mis hõlmaks vähem sisendandmeid. Selle magistritöö eesmärk oli uurida, millised andmed on residentsuse määramiseks kõige olulisemad ja kuidas saavad elanikkonna määramisega vähendatud andmete kontekstis hakkama masinõppe meetodid. Töö uurimisküsimused olid:

1. Millised tunnused praegu olemasolevatest andmetest on kõige vajalikumad, et määrata residentsust?
2. Kuidas toimivad vähendatud andmestikul masinõppe algoritmid?

Töö tulemustest selgus, et olulisemad tunnused, mille abil residentsust määrata, saab jaotada laias laastus kolmeks: eelnevate aastate tulemusega seotud, rahvastikuregistriga seotud ja olulisemate elumärkidega seotud. Täpsemalt leiti peakomponentide analüüsi abil 30 tunnust. Edaspidi saaks katsetada veel erinevaid meetodeid ja hinnata mudelite tulemust üksikuid tunnuseid eemaldades/lisades. Näiteks võiks vaadata, millised tulemused tuleksid kasutades rohkemat kui kahte peakomponenti. Põhiprintsiip võiks olla, et kasutusel oleks nii palju tunnuseid kui vajalik, et maksimeerida kvaliteeti, kuid nii vähe kui võimalik, et lihtsustada ja kiirendada alaliste elanike kogumi kokkupanemist. Näiteks on ressursisäästlikum hõivata andmeid seitsmest registrist senise kaheksateistkümne asemel. Samas on allikate paljusus ka eelis ning senise, 1. jaanuari alaliste elanike määramiseks saab ka edaspidi kasutada senisel kujul residentsuse indeksit.

Siiski näitasid masinõppe algoritmid, et suudavad kiirelt ja tõhusalt ka vähendatud lähteandmete pealt leida üles residendid ja mitteresidendid. Samas on mudelite treenimine sõltuv varasemast olukorrast, mis eeldab mingi baasteadmise olemasolu, mille pealt mudeleid treenida. See on piirav tegur ning seab mudelite toimimise sõltuvusse eelnevatest tulemustest. See omakorda muudab oluliseks tulemuste verifitseerimise ka näiteks valikuuringute abil.

Selle töö põhijärelduseks ja põhipanuseks on, et alaliste elanike määramine võrreldaval tasemel senise metoodikaga on võimalik ka väiksema tunnuste hulgaga. Samas on oluline hinnata selle metoodika paikapidavust ka reaalajas, proovides määrata alaliste elanike arvu näiteks aasta keskel ning hiljem võrreldes seda tulemust senise metoodika pakutava tulemusega. Selleks on hea aeg uue määruse rakendumiseni jääv aeg. Euroopa Komisjoni uus määrus hakkab kehtima 2026. aastast, seega on metoodika arendamiseks ja testimiseks veel mõni aasta aega.

Viidatud kirjandus

Bakker, Bart F. M., Rooijen, Johan van ja Toor, Leo van (2014). "The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics". *Statistical Journal of the IAOS* 30(4). Publisher: IOS Press, lk. 411–424. DOI: 10.3233/SJI-140803.

Chen, Tianqi *et al.* (2023). *xgboost: Extreme Gradient Boosting*. Versioon 1.7.5.1.

CROS (8. mai 2019a). *Administrative register*. CROS - European Commission. URL: https://ec.europa.eu/eurostat/cros/content/administrative-register_en (vaadatud 19.02.2023).

CROS (8. mai 2019b). *Coverage*. CROS - European Commission. URL: https://ec.europa.eu/eurostat/cros/content/coverage-0_en (vaadatud 19.02.2023).

CROS (9. mai 2019c). *Target population*. CROS - European Commission. URL: https://ec.europa.eu/eurostat/cros/content/target-population_en (vaadatud 19.02.2023).

ELT (2013). "Euroopa Parlamendi ja nõukogu määrus (EL) nr 1260/2013, 20. november 2013, Euroopa rahvastikustatistika kohta". *Euroopa Liidu Teataja* L 330, lk. 39–43.

Euroopa Komisjon, toim. (1995). *Statistics on persons in Denmark: a register-based statistical system*. Luxembourg: Office for Official Publ. of the European Communities.

Euroopa Komisjon (2021). *Andmete kogumine – Euroopa rahvastikustatistika*. Andmete kogumine – Euroopa rahvastikustatistika. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12958-Andmete-kogumine-Euroopa-rahvastikustatistika_et (vaadatud 19.02.2023).

Eurostat (2015). *Demographic statistics: a review of definitions and methods of collection in 44 European countries: 2015 edition.* LU: European Commission. Statistical Office of the European Union.

Eurostat (2023). *Population (national level) (demo_pop)*. URL: https://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm (vaadatud 24.02.2023).

Grajalez, Carlos Gómez *et al.* (2013). "Great moments in statistics". *Significance* 10(6), lk. 21–28. DOI: 10.1111/j.1740-9713.2013.00706.x.

Hearty, John (2016). *Advanced Machine Learning with Python: Solve Challenging Data Science Problems by Mastering Cutting-Edge Machine Learning Techniques in Python*. Birmingham, United Kingdom: Packt Publishing Limited.

Kuhn, Max (2008). "Building Predictive Models in *R* Using the **caret** Package". *Journal of Statistical Software* 28(5). DOI: 10.18637/jss.v028.i05.

Lange, Anita (2014). "The population and housing census in a register based statistical system". *Statistical Journal of the IAOS* 30(1), lk. 41–45. DOI: 10.3233/SJI-140798.

Lantz, Brett (2019). *Machine Learning with R: Expert techniques for predictive modeling*. 3. väljaanne. Packt Publishing Limited.

Liaw, Andy ja Wiener, Matthew (2002). "Classification and Regression by randomForest". *R News* 2(3), lk. 18–22.

Lothian, Jack, Holmberg, Anders ja Seyb, Allyson (2019). "An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics". *Journal of Official Statistics* 35, lk. 137–165. DOI: 10.2478/jos-2019-0007.

Maasing, Ethel (2015). "Eesti alaliste elanike määratlemine registripõhises loenduses".

Maasing, Ethel ja Tiit, Ene-Margit (2016). *Implementation of the residency index in demographic statistics*. URL: https://www.stat.ee/sites/default/files/2021-01/Implementation%20of%20the%20residency%20index%20in%20demographic%20statistics.pdf (vaadatud 01.10.2022).

Maasing, Ethel, Tiit, Ene-Margit ja Vähi, Mare (2017). "Residency index – a tool for measuring the population size". *Acta et Commentationes Universitatis Tartuensis de Mathematica* 21(1), lk. 129–139. DOI: 10.12697/ACUTM.2017.21.09.

Marsland, Stephen (2014). *Machine Learning: An Algorithmic Perspective*. 2. väljaanne. Chapman ja Hall/CRC. DOI: 10.1201/b17476.

Meres, Koit (2017). "Rahvaarvu arvutamine: residentsuse indeks vs. rahvastikuregister". *Eesti Statistika Kvartalikiri* (1), lk. 57–66.

Meyer, David *et al.* (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. Versioon 1.7-11.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, United States: MIT Press.

OECD (2017). *Data quality*. Paris: OECD, lk. 77–83. DOI: 10.1787/9789264279889-9-en.

OpenAI (2023). *GPT-3.5*. URL: https://openai.com/ (vaadatud 09.05.2023).

Poston, Dudley L., toim. (2019). *Handbook of Population*. Handbooks of Sociology and Social Research. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-10910-3.

Rahandusministeerium (2022). *Riiklik statistika*. URL: https://www.fin.ee/riik-ja-omavalitsused-planeeringud/riigihaldus/riiklik-statistika (vaadatud 19.02.2023).

Robin, Xavier *et al.* (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". *BMC Bioinformatics* 12, lk. 77.

RT (2022). "Riikliku statistika seadus". *Riigi Teataja* RT I, 11.03.2022, 2.

Siseministeerium (2022). *Rahvastikuregister*. URL: https://siseministeerium.ee/tegevusvaldkonnad / rahvastikutoimingud / rahvastikuregister (vaadatud 19.02.2023).

Statistikaamet (2022). *2021. aasta registripõhise loenduse metoodika kirjeldus*. URL: https://www.stat.ee/sites/default/files/2022-06/Registrip%C3%B5hise%20loenduse%20metoodika%20raport.pdf.

Statistikaamet (2023). *Konfidentsiaalsete andmete kasutamine teaduslikul eesmärgil*. URL: https://www.stat.ee/et/avasta-statistikat/kusi-statistikat/konfidentsiaalsete-andmete-kasutamine-teaduslikul-eesmargil (vaadatud 08.05.2023).

Therneau, Terry ja Atkinson, Beth (2022). *rpart: Recursive Partitioning and Regression Trees*. Versioon 4.1.16.

Valner, Sulev (2012). "Kui palju meid siis ikkagi on?" *Postimees*.

Wallgren, Anders ja Wallgren, Britt (2020). "Comments on the scientific basis of the register-based census". *Statistical Journal of the IAOS* 36(4), lk. 1295–1297. DOI: 10.3233/SJI-200736.

Wallgren, Anders ja Wallgren, Britt (2022). *Register-based statistics: registers and the national statistical system*. 3. väljaanne. Hoboken, NJ: John Wiley & Sons.

Wickham, Hadley *et al.* (2019). "Welcome to the Tidyverse". *Journal of Open Source Software* 4(43), lk. 1686. DOI: 10.21105/joss.01686.

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Egle Saks**,
*(*autori nimi)

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

   Eesti alaliste elanike määramine kasutades masinõppe meetodeid,
   *(*lõputöö pealkiri)

   mille juhendaja(d) on Terje Trasberg ja Raivo Kolde,
   *(*juhendaja nimi)

   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.

Egle Saks
09.05.2023