TARTU ÜLIKOOL

Arvutiteaduse instituut

Informaatika õppekava

Agnes Annilo

Krooniliste astmahaigete pikatoimeliste ravimiandmete imputeerimine

Bakalaureusetöö (9EAP)

Juhendaja: PhD Raivo Kolde

Tartu 2022

Sisukord

Krooniliste astmahaigete pikatoimeliste ravimiandmete imputeerimine

Lühikokkuvõte:

Imputeerimine on statistiline meetod puuduolevate väärtuste asendamiseks andmestikus. Puuduolevate andmete tõttu võib analüüs ning terve uuring olla vigane ja mitte peegeldada üldkogumit. Selleks, et tulevased analüüsid ning mudelid, mis aitavad arstidel patsiente paremini ravida, oleksid täpsemad, on vajalik, et puuduvad andmed asendatakse võimalikult tõepäraselt. Bakalaureusetöö eesmärk oli uurida võimalikke imputeerimise meetodeid Eesti astma- ning kroonilise obstruktiivse kopsuhaigusega patsientide andmetel ning viia läbi nende võrdlus. Töö teoreetilises osas kirjeldatakse astma ravimeetodeid ja imputeerimismeetodite ning analüüsimeetodite teoreetilist tausta. Lõpus tehakse rakendatud imputeerimismeetodite analüüs ning antakse autori hinnang meetodite rakendatavusele.

CERCS teaduseriala: P160 statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Võtmesõnad: Imputeerimine, *hot deck*, kNN, otsustusmets

Imputing Long Acting Medication Data of Chronic Asthma Patients

Abstract:

Imputation is a statistical method to handle missing data. Missing data jeopardizes the ability of an analysis to accurately reflect the parent population. It is necessary that missing data are replaced correctly, to ensure that future analyses and models, which help medical specialists treat patients are more precise. The aim of this Bachelor's thesis was to study different imputation methods using the data of Estonian asthma and chronic obstructive pulmonary disease patients and to conduct the comparison of said methods. The theoretical part of this thesis describes the possible ways to treat asthma and gives an overview of imputation and the possibility to compare imputation methods. Finally, an analysis of the chosen imputation methods and the author's assessment of these methods is given.

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

Key words: Imputation, hot deck, kNN, random forest

Sissejuhatus

Andmete kvaliteedi üheks tingimuseks on andmete terviklikkus, see tähendab, et mida enam on andmestikust andmeid puudu, seda ebakvaliteetsemad on andmestiku põhjal tehtud analüüsid. Kõige efektiivsemate ning kiiremate imputeerimise meetodite väljaselgitamine aitaks teha Eesti digiandmetelt loodud andmestikke analüütikute jaoks kvaliteetsemaks ning analüüsitulemusi täpsemaks. Kõige olulisem on leida meetodeid, mis sobivad andmestiku struktuuriga, st olemasolevate tunnustega. Eestis on digiretseptide sisestus standardiseeritud, seega antud andmestikul saadud leiud võiksid olla üldiselt rakendatavad ka teistele ravimigruppidele. Kvaliteetse ning kindla imputeerimismeetodi leidmine on vajalik eeltöö, et sarnaste andmestike analüüsil oleks võimalik teha otsus, milliseid imputeerimismeetodeid eelistada.

Bakalaureusetöös on kasutatud Eesti retseptide digiandmeid, spetsiifiliselt astma ning kroonilise obstruktiivse kopsuhaigusega (KOK) patsientide andmeid, samuti Haigekassa raviarveid. Töö sihiks on analüüsida andmestikul erinevaid imputeerimise meetodeid, nende klassifitseerimise headust ning puudujääke. Andmed puhastatakse, leitakse sobivad imputeerimismeetodid ning võrreldakse nende rakendamist. Iga meetodi analüüsis kasutatakse sama andmestikku, kus puuduvad andmed on juhuslikult tekitatud. Seejärel võrreldakse meetodeid omavahel nii kiiruses kui ka headuses. Analüüs viiakse läbi kasutades tarkvara R.

1. Terviseandmete imputeerimine

Bakalaureusetöö keskendub Eesti digiretsepti andmetele. See andmestik katab kõik Eestis väljakirjutatud ravimid ning on seetõttu väärtuslikuks allikaks erinevate epidemioloogiliste uuringute läbiviimiseks.

1.1. Digiretsepti andmed

Järgneva jaotise lõigud on kirjutatud kasutades Astma andmete uuringutaotlust (2020).

Eestis on alates 2010. aastast, kui mindi üle Digiretsepti infrastruktuurile, kogutud kõik andmed retseptide väljakirjutamise ning -ostmise kohta ühte kesksesse infopanka. Teoreetiliselt peaks selline andmete kogumine kergendama ravimikasutamise uuringuid. Praktikas tuleb andmete kasutamiseks teha palju lisatööd. Retseptiandmete sisestus ei ole ühtlustatud ning igal arstil võib olla tekkinud enda andmete esitamise viis. Siiski kasutatakse terviseandmeid, et uurida ravimi kasutamise levimust, ravi muutust ajas, ravi efektiivsust ja ravimisoostumust ning võrrelda tulemusi teiste Euroopa riikidega. Lisaks Digiretseptide andmetele on bakalaureuse-töösse kaasatud ka Haigekassa raviarvete andmed.

Astma ning krooniline obstruktiivne kopsuhaigus ehk KOK on esinduslikud näited kroonilistest haigustest, mis on laialt levinud, mille ravile on selged juhised ning mille ravi eeldab pidevat meditsiinilist jälgimist. Krooniliste ravimite retseptide puuduseks on see, et patsientidele on pikaajalise ravimi võtmise järel annustamine selge ning seetõttu on andmestikus annustamise osa puudulikult täidetud. Andmete analüüsi teevad sellised puuduvad väljad aga keerukamaks (Astma andmete uuringutaotlus, 2020).

Analüüs tehti ATC koodi põhjal grupeeritud andmestikes. ATC kood ehk Anatoomilis-terapeutiline keemiline kood on unikaalne identifikaator, mis kirjeldab ravimi kasutusvaldkonda (European Medicines Agency, 2022). Nii grupeerimata kui ka grupeeritud andmestikes on puudu 85% tunnuse „annustamise päevi" andmetest. See on väga kõrge puuduvate andmete protsent, mille täielikuks imputeerimiseks on ilmselt vaja teha kvalitatiivset uuringut. Bakalaureusetöös kasutati imputeerimiseks vaid andmeid, millel oli tunnuse „annustamise päevi" kõik väärtused olemas. Puuduvate andmete eemaldamise järel genereeriti juhuslikult igasse ATC klassi tunnusele „annustamise päevi" 10% puuduvaid andmeid. Saadud andmestikul viidi läbi kõik töös esitatud analüüsid.

1.2. Puuduvad andmed

Järgnev jaotis on kirjutatud toetudes aine Andmeanalüüs 2 loengukonspektile (i.a), Fieldingi jt artiklile (2008) ning Pälli ja Maiväli õpikule (i.a).

Tihti on suuremates uuringutes mitmeid puuduvaid väärtuseid. Puuduvate andmete tõttu tekivad nihkega hinnangud ning analüüsi tulemused on vigased, eriti kui andmed puuduvad juhuslikult. Valesti hinnatud andmete puudumise mehhanism võib põhjustada kallutatud tulemusi, mistõttu on oluline teha puuduvatele andmetele analüüs või võtta rangeks eelduseks kindel puudumise struktuuri tüüp.

Puuduolevad andmed jaotatakse kolme rühma.

Täiesti juhusliku puudumise ehk *missing completely at random (MCAR)* puhul on puuduvate andmete jaotus juhuslik, ehk väärtuse puudumine ei sõltu väärtusest endast ega ka vaadeldavatest parameetritest uuringus. MCAR määramiseks on ranged eeldused, tänu millele on puudumise struktuuri tüübi määramise otsus kõige kindlam. Täiesti juhuslik puudumine meditsiiniandmestikes võib tekkida näiteks meditsiinilise tehnika rikke või proovide väärkäitlemise tõttu.

Missing at random (MAR) ehk juhusliku puudumise puhul ei ole kadu vaadeldud andmete puhul täiesti juhuslik ning võib sõltuda mingi teise tunnuse väärtusest. Väärtuse kadu võib olla sõltuv teistest parameetritest, kuid mitte puuduvast väärtusest endast. Selline andmekadu võib tekkida, kui mõnes uuringus ei tehta teatud vanusegrupil mingisuguse tunnuse mõõtmisi, näiteks südame rütmi, kuna tavaliselt ei ole vanusegrupil sellega probleeme. Siin on tunnuse puudumine mõjutatud mingist teisest tunnusest (vanusest), kuid ei ole sõltuv (eeldatavatest) mõõtmistulemustest. Kõige edukamalt saab statistilisi metoodikaid rakendada andmetele, mis on MAR.

M*issing not at random (MNAR)* ehk mittejuhuslikult puuduvate andmete puhul sõltub puudumine puuduolevast väärtusest endast. Üheks näiteks on küsitlus, milles on küsimused, millele teatud osa üldkogumist ei soovi vastata. Näiteks on võimalik, et alla 18-aastased suitsetajad ning alkoholitarbijad ei soovi küsimustikus anda vastust küsimustele suitsetamise või alkoholi tarbimise kohta. Siin on mittevastamise põhjuseks puuduvad väljad ise, s.t on tõenäoline, et vastamata jätnud noored ei soovinud vastata, kuna nad suitsetavad või tarbivad alkoholi. Andmed on MNAR, kui nad ei ole MCAR ega MAR. On võimalik, et ühes

andmestikus leidub igat tüüpi andmekadu (Fielding, jt, 2008; Päll ja Maiväli, i.a; Andmeanalüüs 2 loengukonspekt).

1.3. Imputeerimine

Puuduvate väärtustega objekte on võimalik analüüsi kaasata, eemaldades või kompenseerides puuduvad väärtused. Kõige lihtsamaks ning laialdasemaks lahenduseks on ridade eemaldamine, mille kasutamise eelduseks on vähene ning juhuslik andmekadu. Selle meetodi puhul eemaldatakse read, millel on üks või mitu puuduvat väärtust, kuid erinevate puuduvate väärtuste puhul võib tekkida suur kadu (Zhang, 2016). Töös kasutatavas andmestikus olev 85% kadu liialt suur, et ridade eemaldamise meetodit tervele andmestikule rakendada. Seevastu eemaldati andmestikust loogikavigadega kirjed.

Üks võimalikest viisidest kadu kompenseerida on imputeerimine ehk asendusväärtuste leidmine andmelünkade täitmiseks. Imputeerimise puhul asendatakse puuduolev väärtus selle hinnanguga. Dondes jt (2006) märgivad, et imputeerimise eesmärk ei ole saada puuduvatele väärtustele kõige täpsemat hinnangut, vaid saada väärtuste kombinatsioon, mis võiks esineda ka üldkogumis. Imputeerimise meetodid põhinevad statistilisel printsiibil, et uuringu tulemus ei tohiks sõltuda valimist ning tulemus peaks kehtima ka uue valimiga. Seetõttu võib iga objekti juhuslikult valitud valimist asendada teise juhuslikult üldkogumist valitud objektiga. Sellist asendamist püüavadki imputeerimise meetodid matkida. Siiski suuremahulise kao puhul ei pruugi klassikalised imputeerimise meetodid töötada (Dondes jt, 2006).

Imputeerimise meetoditel on üldiselt sarnased eeldused. Van Buuren (i.a) märgib, et peamised eeldused on seotud puuduvate andmete tüübiga ning eelistatakse andmete juhuslikku puudumist. Kvaliteetse imputeerimise eelduseks on hea ülevaade andmetest, et määrata korrektselt puuduvate andmete tüüp. Töös kasutatud andmete puhul on tegemist juhusliku puudumisega. Annustamise päevade arvu kirjutamata jätmine ei ole tingitud sellest, mitmeks päevaks ravimit jätkub, vaid näiteks ravaimist või patsiendist endast (on võimalik, et talle on ravimit varem välja kirjutatud). Imputeerimise metoodika oleneb andmete tüübist, üldiselt on võimalik imputeerida arvulisi või kategoorilisi, harvem nominaalseid tunnuseid. Samuti on erinevatel meetoditel eeldused imputeeritava tunnuse jaotuse kohta (van Buuren, i.a).

Bakalaureusetöö eesmärk on leida võimalikult hea imputeerimise meetod või meetodid, mis ei eelda eelnevat andmete kvalitatiivset analüüsi ning mida on lihtne puhastatud andmetel

rakendada. Andmeid on tarvis imputeerida, kuna see võimaldab läbi viia vajalikke teadusuuringuid nii astma retseptidel kui ka teiste haiguste retseptidel.

2. Andmete kirjeldus

Terviseandmete kvaliteetseks imputeerimiseks on vajalik omada ülevaadet andmestikus olevatest andmetest ning nende meditsiinilisest taustast.

2.1. Astma ja krooniline obstruktiivne kopsuhaigus ning nende ravi

Astma on põletikuline hingamisteede krooniline haigus, mille haigushooga kaasneb hingamisteede õhuvoolu takistus. Haigushoo sümptomid on hingeldamine, köha ning pingetunne rinnus. Sümptomeid on võimalik leevendada pikemaajalise püsiravi ehk baasravi või lühiajalise hooravi abil (Ravijuhendite Nõukoda, 2020). Eestis on ligikaudu viiel kuni kaheksal protsendil elanikkonnast diagnoositud astma (Haigekassa, 2017).

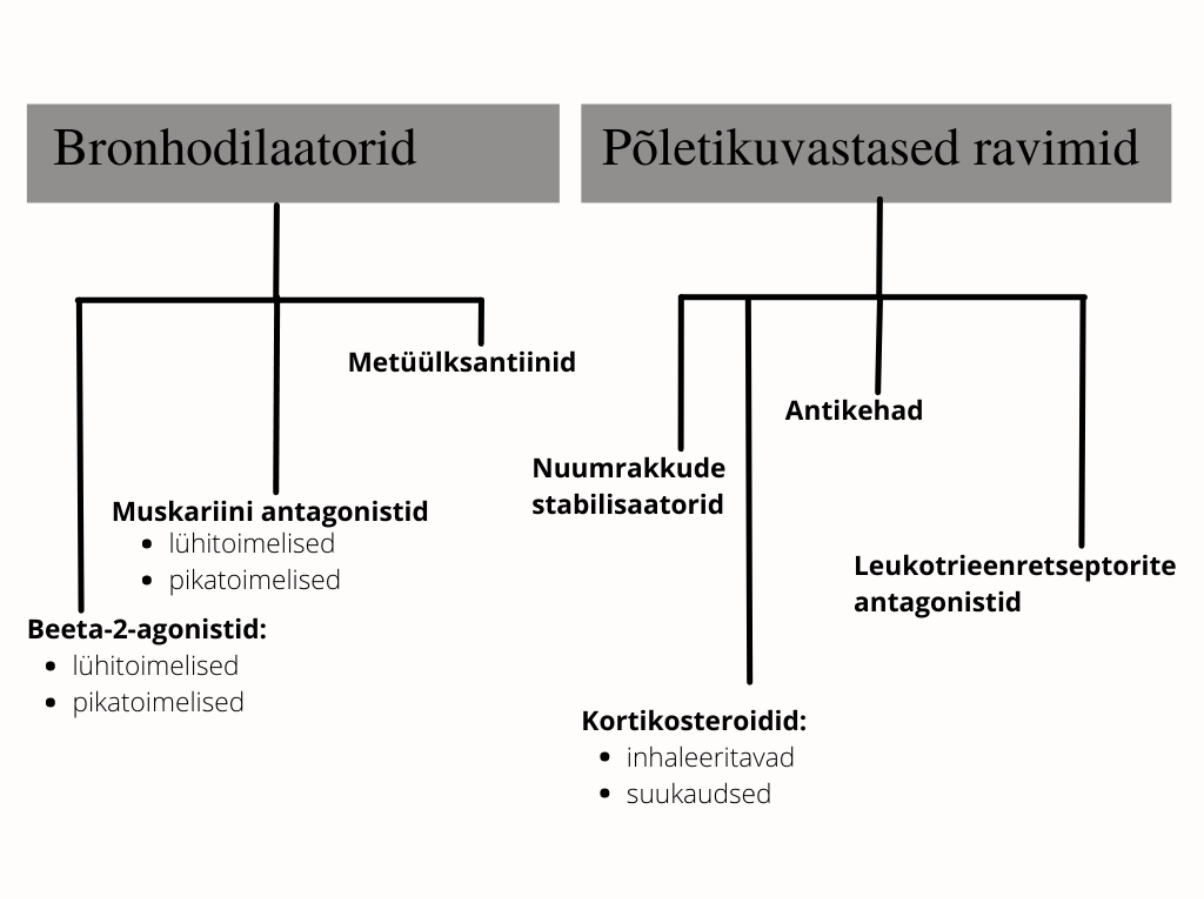
Kroonilise obstruktiivse kopsuhaiguse ehk KOK-i ravi sarnaneb astma ravile, kuid peamiselt kasutatakse bronhodilaatoreid muskariini antagonistidega. KOK ravimid jagunevad samuti nii pika- kui ka lühitoimeliseks, kuid KOK diagnoosiga patsiendid peavad üldjuhul võtma ravimeid igapäevaselt (Eesti Kopsuliit, 2022).

Eesti Ravijuhendite Nõukoja ravijuhendis (2020) on kirjeldatud nii hooravi kui ka baasravi. Astmahoogu kiirelt leevendav hooravi kestab üldiselt neli kuni kuus tundi ning ei ole mõeldud pidevalt korduvaks kasutamiseks. Lühiajaliste ravimite pidev kasutamine võib kaasa tuua tüsistusi, näiteks südame rütmihäireid. Pikemaajaline baasravi ennetab astma ägenemist ja vähendab sümptomeid, vähendades bronhide limaskesta turset. Selleks võetakse ravimeid regulaarselt arsti jälgimisel (Ravijuhendite Nõukoda, 2020).

Pikatoimeliste ravimite all peetakse silmas astma profülaktika ravimeid, lühitoimeliste all astmahoo leevendamise ravimeid. Astmahoo leevendamise ravimeid ei võta patsiendid regulaarselt ning ravimi kestus sõltub paljugi päevasest annusest ning astmahoo tõsidusest, seetõttu on keerulisem nende annustamise päevade arvu hinnata (Ravijuhendite Nõukoda, 2020). Bakalaureusetöös analüüsitakse vaid pikatoimelisi ravimeid.

Eduka astma püsiravi järel puuduvad patsiendil päevased ning öised sümptomid ning igapäevategevused on minimaalselt häiritud (Haigekassa, 2017). Astma püsiravi eesmärk on saavutada kontroll astma üle, kuid seda võimalikult väikese annusega. Ravile lähenetakse astmeliselt: alustatakse võimalikult madalate annustega ning arst reguleerib annust vastavalt patsiendi seisundi hinnangule (Ravijuhendite Nõukoda, 2020). Üldiselt jagunevad

astmaravimid kaheks: bronhodilaatoriteks ning põletikuvastaste toimeainetega ravimiteks (joonis 1). *Global Initiative For Asthma* koduleht (2021) märgib, et raviks kirjutatakse enamasti välja mõlemat korraga, ägeda astmahoo kiireks leevenduseks bronhodilaatoreid ning pikemaajalise seisundi kontrollimiseks põletikuvastaseid ravimeid. Leidub kombineeritud ravimeid, kuid kuna kombineeritud ravimit tuleb kasutada regulaarselt ühtlase annusena, võib juhtuda, et ägeda astmahoo korral on lühiajalist ravimit liiga vähe. Selle ohu vältimiseks võib patsientidel olla lisaks ka retsept bronhodilaatorile (Global Initiative for Asthma, 2021).



Joonis 1. Astmaravimite klassifikatsioon (Kiboneka, 2020)

Ravijuhendite Nõukoda (2020) annab ravijuhendis ülevaate eelistatud astma ravi metoodikast. Esmase diagnoosi saanud patsiendid saavad tavaliselt ravimiks lühitoimelisi beeta-2-agoniste, mida kutsutakse enamasti lühendiga SABA (*Short Acting Beta Agonists*) kombineerituna kortikosteroididega. Ravi soovitud tulemuse mittesaavutamise korral kirjutatakse välja pikatoimelisi beeta-2-agoniste, tihti kasutatakse neile viitamisel lühendit LABA (*Long Acting Beta Agonists*), mis sobivad raviks ainult kombineerituna inhaleeritava glükokortikosteroidiga (Ravijuhendite Nõukoda, 2020).

2.2. Andmestik

Töös kasutatud andmed on Eesti digitaalsed terviseandmed. Valimi moodustavad isikud, kellel oli perioodil 01.01.2011 - 31.12.2019 vähemalt üks KOK (kroonilise obstruktiivse kopsuhaiguse) või astma diagnoos. Andmestikus on ligikaudu kolm miljonit väljakirjutatud retsepti.

Andmestikus on järgmised väljad: „isiku ID", „sugu", „sünnikuupäev", „retsepti väljakirjutamise kuupäev", „retsepti väljaostmise kuupäev", „retsepti toimeaine ATC-kood", „annuse kogus" ehk kui suur on üks annus, „annustamise kordi päevas" ehk mitu korda päevas ravimit peab manustama, „annustamise päevi" ehk retseptist patsiendile jätkumise päevade arv. Bakalaureusetöö eesmärk on imputeerida puuduolevad väärtused tunnusest „annustamise päevi".

Lisaks kasutati analüüsiks juurde arvutatud väljasid haigekassa raviarvete andmestiku põhjal. Haigekassa raviarvete andmestikus on igale patsiendile määratud diagnoos, mille põhjal neile ravim hüvitatakse. Selle andmestiku väljad on isiku ID, arve alguskuupäev, arve lõppkuupäev, diagnoosid.

Kõiki välju ei ole imputeerimiseks vaja - ilmselt ei mõjuta retsepti annustamise päevade arvu retsepti väljakirjutamise või väljaostmise kuupäev. Samuti on keeruline kaasata tunnust annustamise kordi päevas, kuna ka seal on palju puuduvaid väärtuseid.

Lõplikusse analüüsi on kaasatud järgmised tunnused:

- isiku id
- vanus retsepti välja kirjutamise hetkel (arvutatud kasutades sünnikuupäeva ning väljakirjutamise kuupäeva)
- sugu
- annustamise päevade arv (imputeeritav tunnus)

Lisatunnused, mis on teiste tunnuste pealt leitud:

- diagnooside koodid (binaarsete tunnustena: 1- patsiendil on diagnoos, 0 - ei ole diagnoosi):
    - J44 - Muu krooniline obstruktiivne kopsuhaigus
    - J45 - Astma
    - J46 - Astmaatiline seisund (Sotsiaalministeerium, i.a)

- korduvretseptide arv
- isikule väljakirjutatud retseptide arv, ATC koodi järgi grupeerituna
- iga kirje väljakirjutamise indeks, ehk mitu korda oli isikule eelnevalt sama ravimit välja kirjutatud

Diagnoosid lisati lisatunnusena eeldusega, et erinevate diagnooside kombinatsioonidega isikutele kirjutatakse ravimeid erinevalt välja. Korduvretseptide arvu võeti arvesse, kuna ravimeid kirjutatakse korduvretseptidena ilmselt sarnaselt välja. Isikule välja kirjutatud retseptide arvu peeti oluliseks lisatunnuseks, kuna see näitab, kui pikalt patsient on ravimit võtnud. Seda, mitu korda ravimit eelnevalt välja kirjutati võeti kasutusele, et kirjeldada ravi muutust ajas ilma, et peaks kasutama kuupäevade tunnuseid.

Andmete puhastamiseks kasutati osaliselt Kaari Kuusi eelmise aasta bakalaureusetöö koodi (Kuus, 2021). Selle käigus tuli andmestikust eemaldada loogilised ja jämedad vead, näiteks asendada vigased ATC koodid tegelike koodidega.

Käesolevas töös imputeeritakse tunnust “annustamise päevade arv”. Annustamise päevade arvu täitmine retsepti kirjutamisel ei ole kohustuslik. Siiski on täitmise puhul ette nähtud, et märgitud oleks toimeaine täpne sisaldus ravimis ning väljastavate ühikute koguarv või ühekordne annus, annustamise sagedus ning ravi kestus (Riigi Teataja, 2005). Sellest hoolimata on retseptidesse märgitud info üldiselt puudulik.

2.3. Andmestiku puhastamine

Andmestikus on nii pikaajalised kui ka lühitoimelised retseptid, millest imputeerimisel kasutatakse vaid pikatoimelisi ravimeid. Kroonilise haiguse nagu astma või KOKi ravi puhul tuleb ravimit välja kirjutada enam kui kahekuuliseks raviks vajaminev kogus, erandiks on ravi alustamine või muutmine. Pikatoimelisi ravimeid kirjutatakse seetõttu välja korduvretseptina ehk kahe- või kolmekordse retseptina (Riigi Teataja, 2005). Patsiendile kirjutatakse samal kuupäeval välja mitu sama toimeainega ravimit, mille puhul annustamise päevade arv kehtib summaarselt ravimitele, näiteks kestab üks retsept kolmekordsest korduvretseptist kolmandiku „annustamise päevi” tunnusest. Selleks, et see tunnus kajastuks andmetes õigesti, tuli korduvretseptide andmed puhastada.

Andmestikus leidus juhuseid, kus patsiendile oli samal päeval kirjutatud välja üle 20 identse retsepti. Samuti leidus patsiente, kellele oli samal päeval kirjutatud välja sama ATC koodiga, kuid erinevate annustamise päevadega retsepte. Korduvretseptidena identifitseeritakse vaid täiesti identsed retseptid (välja arvatud ravimi väljaostmise kuupäev). Ka siis leidus üle kolme korraga kirjutatud retsepti. Kui korduvaid retsepte oli päevas üle 3, arvestati vaid nende kirjetega, millel oli väljaostmise kuupäev olemas, muid loeti sisestusvigadeks.

Selleks, et olla imputeerimisel võimalikult robustne ning hoiduda sisestusvigadest ja kõrvalekalletest, kaasati analüüsi vaid need retseptid, mille annustamise päevade arvu sagedus oli andmestikus üle 1000. Kirjed, kus annustamise päevade arv oli vähem või võrdne viiega, grupeeriti üheks väärtuseks Mitmel ravimil puudus igal kirjel tunnuse „annustamise päevi" väärtus, mistõttu on lõplikusse analüüsi kaasatud vaid ravimid, mille puhul leidus üle 100 retsepti, kus annustamise päevade arv oli täidetud. Ravimid, mille annustamise päevade arv oli null, asendati *NA*-ga (*not available)*, mis R-koodis viitab puuduvale väärtusele. Imputeerimismeetodite rakendamiseks eemaldati andmestikust kõik kirjed, kus tunnus „annustamise päevi" oli puudu, mida oli andmestikus 906 851 kirjel 1 055 918 kirjest. Annustamise päevade arv muudeti 14 klassiga kategooriliseks tunnuseks.

3. Andmestiku imputeerimine

Antud peatükis antakse ülevaade andmete imputeerimisest ning levinud metoodikast. Lisaks antakse ülevaade võimalikest imputeerimismeetodite kvaliteedi võrdluse meetoditest. Lõpuks kirjeldatakse lühidalt levinud ning eelistatud mitmest imputatsiooni ning miks seda ei kasutatud.

3.1. Elementaarsed meetodid

Ühekaupa imputatsioon asendab tunnuse kõik puuduvad väärtused ühe hinnanguga. Üheks võimaluseks on asendada puuduvad väärtused olemasolevate väärtuste keskmise, moodi või mediaaniga, täpsemate tulemuste jaoks arvestatakse mõne teise tunnuse väärtustega. Need meetodid põhjustavad dispersiooni alahinnangut ning kõik analüüsi tulemused on seetõttu nihkega. Seetõttu kasutatakse meetodeid vaid väga väikese kao puhul (Zhang, 2016). Kuna töös tekitatakse puuduvaid andmeid vaid 10%, kasutati imputeerimiseks ka moodiga asendamist. Moodiga asendamiseks leiti igast ravimigrupist tunnuse „annustamise päevi“ mood. Moodiga imputeeritud andmete täpsust kasutati teiste meetodite täpsuse võrdlusena. Kuna tegelikus andmestikus on puudu 85% andmetest, otsustati moodiga imputeerimist mitte meetoditevahelisse võrdlusesse kaasata.

Lõputöös on imputeerimiseks kasutatud segaandmeid: nii pidevaid, diskreetseid kui ka kategoorilisi tunnuseid. Klassifitseeritav tunnus on kategooriline. Selline andmete segu kitsendab oluliselt võimalikke imputeerimismeetodeid. Antud kitsenduste põhjal on valitud kolm lihtsasti rakendatavat ning üldiselt kiiret meetodit, mida võiks antud andmestiku imputeerimisel kasutada.

3.2. *Hot deck*

Jaotis on kirjutatud toetudes Kowariki ning Templi (2016) ja Andridge ning Little (2010) artiklitele.

Hot deck metoodika on pärit aegadest, kus andmeid hoiti perfokaartidel, kuid on siiani populaarne tänu oma lihtsusele ning kiirusele. *Hot deck* meetodite puhul kasutatakse imputeerimiseks sama andmestiku väärtuseid. Andmestikust leitakse puuduoleva väärtusega kirjele võimalikult sarnane kirje, millel on imputeeritav väärtus olemas. Sellist kirjet kutsutakse doonoriks, kuna tema olemasoleva väärtusega asendatakse puuduolev väärtus. Mõnel juhul

valitakse doonor mitme võimaliku doonori seast (juhuslikud *hot deck* meetodid), teistel juhtudel valitakse vaid üks "lähim naaber", mille tunnuse väärtust kasutatakse (deterministlikud *hot deck* meetodid). Probleemiks on juhud, kui mitmel imputeeritaval väärtusel on sama doonor, kuna see vähendab väiksemate klasside esinemissagedusi. Seetõttu on mõnel meetodil iga doonori kohta piirang, mitu korda seda kasutada lubatakse. Doonori kasutamise arvu piiramine mõjutab seevastu võimalust valida kõige parem ning lähedasem doonor.

Bakalaureusetöös kasutatud *hot deck* meetod paketist *VIM* on järjestikune ning juhuslik imputatsioonialgoritm (Templ, jt, 2021). Algoritmi idee seisneb puuduoleva väärtusega kirjele võimalikult sarnase kirje või n.ö doonori leidmises. Tunnus, kus esialgsel kirjel puudub väärtus, asendatakse doonori sama tunnuse olemasoleva väärtusega. Järjestikuses ning juhuslikus *hot deck* meetodis leitakse doonor, pannes esiteks paika abitunnused, mille järgi doonor leitakse. Abitunnused valib meetodi rakendaja. Abitunnuste põhjal järjestatakse andmestik ning käiakse see algusest lõpuni läbi. Doonoriks võetakse esimene kirje, millel on puuduoleva tunnuse väärtus olemas. Igat järgnevat kirjet võrreldakse seejärel doonoriga ning kui kirjel ei ole puuduvat väärtust, valitakse hoopis see doonoriks. Kui kirjes on puuduolev väärtus, asendatakse see viimatise doonori olemasoleva väärtusega (Kowarik ja Templ, 2016; Andridge ja Little, 2010).

3.3. K- lähimat naabrit

Jaotis toetub Kowariki ning Templi artiklile (2016) ja Eurostat käsiraamatule (2014).

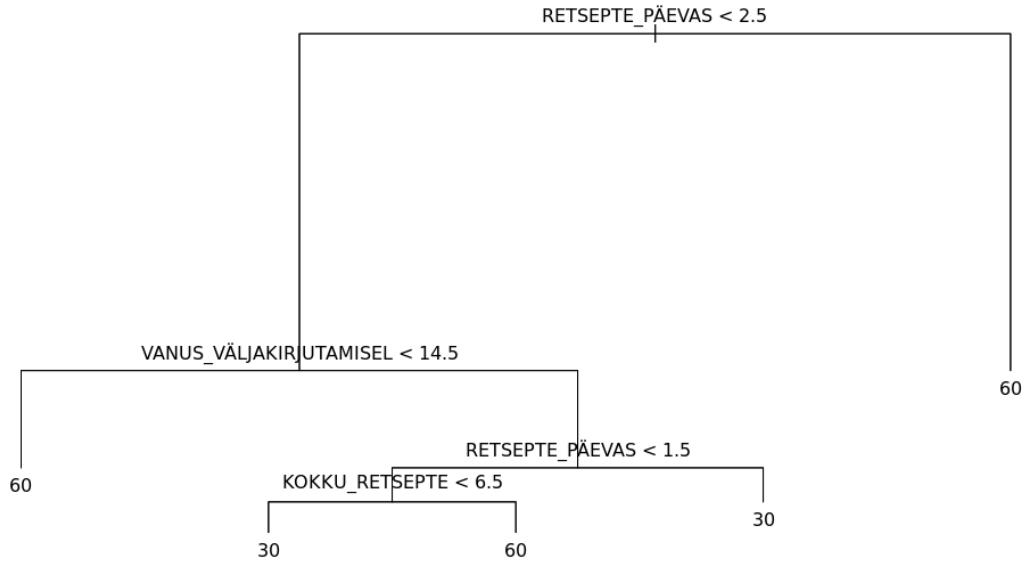
Hot deck metoodikale on sarnane *k* lähima naabri metoodika (kNN). KNN algoritmi idee on leida *k* kõige lähimat naabrit kirjele *i*, kasutades kaugusemõõdikut D, ehk on vaja, et *D(i,k)* oleks minimaalne. Väärtuse imputeerimiseks leitakse *k* doonori imputeeritava tunnuse väärtuse agregaat, mis oleneb tunnuse tüübist. Kategooriliste tunnuste puhul kasutatakse moodi ning kui kahe väärtuse sagedus on võrdne, valitakse väärtus nende vahel juhuslikult.

KNN algoritmi rakendamiseks on esiteks vaja panna paika muutuja k väärtus. Arv *k* väljendab algoritmis mitut kategoriseerimist vajava punkti naabrit vaadatakse. Paketi *VIM* kNN meetodis on vaikimisi k väärtus 5, mida peetakse piisavaks k väärtuseks (Templ, et al., 2021). Lähimad naabrid leitakse mingisuguse vahemaa mõõdikuga, bakalaureusetöös Goweri kauguse edasiarendusega. Goweri kaugus kahe vaatluse vahel on kaalutud keskmine igast tunnusest

saadud sisendist, kus kaal esindab tunnuse olulisust. Iga tunnuse sisend arvutatakse erinevalt pidevatele, binaarsetele ning mitme klassiga kategoorilistele tunnustele. Imputeerimiseks leitakse doonorväärtuste mood. Kui erinevaid väärtuseid on võrdselt, valitakse imputeeritav väärtus juhuslikult (Kowarik ja Templ, 2016; Eurostat, 2014).

3.4. Otsustusmets

Otsustusmets on kombinatsioon mitmest otsustuspuust. Kategoorilistel andmetel põhinev puu klassifitseerib etteantud tunnuse (töös imputeeritava tunnuse) erinevate tingimuste abil. Arvulistele tunnustele antakse tingimused $X < a$ ning $X > a$ ehk, kui $X < a$ liigutakse vasakusse puu harusse (Ripley, 2021). Joonisel 2 on toodud lihtne näide otsustuspuu struktuurist.



Joonis 2. Lihtne näide otsustuspuu struktuurist

Fürnkranz (2010) kirjeldab otsustuspuude struktuuri: tunnuse klassifikatsioon algab puu juurtipust, kus esitatakse esimene klassifitseerimisküsimus (teiste tunnuste põhjal). Küsimuste põhjal klassifitseerimine jätkub, kuni jõutakse lehttipuni, mille väärtus on kategoriseeritava tunnuse väärtus. Otsustuspuud õpivad juurtipust lehttippudeni rekursiivselt „jaga ja valitse" algoritmi järgides. „Jaga ja valitse" algoritm töötab otsustuspuudel, jagades iga sõlme kaheks või enamaks sõlmeks, mida omakorda jaotatakse sõlmedeks, kuni kas sõlme ei ole enam

võimalik jagada või kõik tunnuse väärtused on samast klassist. Algoritm valib esimeseks sõlmeks parima tunnuse, mille väärtused jaotatakse sõlme küsimuste vastustena ehk harudena. Erinevatel algoritmidel on erinevad kriteeriumid, mille järgi pidevaid tunnuseid harudeks jagatakse. Diskreetseid või kategoorilisi tunnuseid on lihtsam jaotada, kuna enamasti valitakse puu haruks üks diskreetne väärtus. (Fürnkranz, 2010)

Bakalaureusetöös on kasutatud juhusliku metsa imputatsioonialgoritmi paketist *missForest* (Stekhoven, 2022). Algoritmi esimene samm on asendada iga puuduv väärtus moodiga. Seejärel järjestatakse veerud nende puuduolevate väärtuste osakaalu põhjal väiksemast suuremani. Peale igat iteratsiooni võrreldakse saadud ning eelneva andmestiku imputeeritud pidevaid ning kategoorilisi tunnuseid ning saadakse tunnuste erinevus. Lõpetamise tingimuseks iga tunnuse erinevuse suurenemine vähemalt ühe korra. Seni, kuni lõpetamise tingimus ei ole täidetud, vaadatakse läbi iga puuduvate väärtustega veerg vastavalt järjestusele. Töös on puuduvate väärtustega veerge vaid üks, seega toimub klassifitseerimine vaid ühel veerul. Sobitatakse juhuslik mets, mis ennustab mittepuuduvad väärtused veerust, kasutades klassifitseerivate tunnustena kõiki ülejäänud tulpasid. Saadud klassifikaatorit kasutatakse, et ennustada tegelikke puuduvaid väärtuseid vaadeldavast veerust. Saadud väärtused imputeeritakse. Edasi kasutab algoritm imputeeritud andmestikku, s.t sobitab juhusliku metsa eelmises iteratsioonis imputeeritud andmestikul. Algoritm töötab mitu korda ning tagastab viimase imputeeritud andmemaatriksi. Selle, mitu iteratsiooni algoritm maksimaalselt läbib, määrab kasutaja. Töös ei ole lõpetamise tingimust määratud ning kasutati meetodi vaikimisi iteratsioonide arvu 10 (Stekhoven, 2022; Stekhoven ja Bühlmann, 2012).

3.5. Mitmene imputatsioon

Järgnev lõik on kirjutatud kasutades Kingi jt artiklit (2001).

Üldiselt eelistatakse ühesele imputatsioonile mitmest imputatsiooni. Mitmese imputatsiooni eesmärk on matkida andmete juhuslikkust mitme erineva, kuid võimaliku imputeeritud andmestiku põhjal. Mitmesed imputatsiooni ehk mitmese asendamise meetodid tekitavad mitu erinevat terviklikku andmestikku. Igas andmestikus on kasutatud samu imputeerimis-meetodeid, vajadusel erinevate parameetritega. Vahel kasutatakse ka erinevaid imputeerimismeetodeid. Imputeerimise järel viiakse igal andmestikul eraldi läbi soovitud analüüs. Iga andmestikku analüüsitakse teistest sõltumatult ning analüüside tulemused

ühendatakse lõpuks üheks punkthinnanguks. Näiteks, kui analüüsi eesmärk on arvutada mingisuguse tunnuse keskmine, imputeeritakse esiteks selle tunnuse väärtused *m* korda. Saadakse *m* andmestikku ning igal andmestikul arvutatakse tunnuse keskmine. Selleks, et saada lõplikku punkthinnangut, kasutatakse valitud agregaatfunktsiooni, üldiselt aritmeetilist keskmist, et leida keskmiste keskmine, mis on lõplik punkthinnang tunnuse keskmisele. Samuti arvutatakse hinnangu hälve ning usaldusintervall. Sarnaselt keskmise leidmisele on võimalik viia läbi ka teistsuguseid analüüse, näiteks mudelite loomist terve imputeeritud andmestiku põhjal (King jt, 2001).

Andmetele, mis ei ole juhuslikult puuduvad, võib mitmene imputatsioon anda valesid tulemusi. Paljud mitmese imputatsiooni meetodid eeldavad imputeeritava tunnuse normaaljaotust. Probleemi vältimiseks on võimalus tunnus normeerida või valida meetod, mis ei eelda normaaljaotust (Sterne jt, 2009). Üldiselt võib mitmene imputatsioon anda täpsemaid tulemusi, kui puudu on palju andmeid, kuid lõppkokkuvõttes sõltub imputatsioonimeetodi valik ja sobivus andmestikust ning puuduvatest andmetest.

Bakalaureusetöö eesmärk ei ole andmestikul läbi viia edasist analüüsi, vaid saada üks imputeeritud andmestik, mistõttu rakendatakse ühest imputatsiooni. Mitmese imputeerimise eeldused on ranged ning meetodid on suurte andmestike peal väga aeglased. Siiski, kui tervet analüüsi viib läbi üks inimene, on mitmene imputatsioon parem ning eelistatavam valik. Ühene imputatsioon omab siiski suurt tähtsust, peamiselt siis, kui imputeeritud andmestik on tarvis edastada teistele spetsialistidele, kes viivad ise edasise analüüsi läbi. Samuti aitab ühese imputatsiooni analüüs anda ülevaadet, milliseid meetodeid on mõttekas mitmeses imputatsioonis kaaluda. Seetõttu, kuigi mitmene imputatsioon võib anda paremaid tulemusi, on oluline kaaluda ka ühest imputeerimist.

3.6. Meetodite võrdlemise meetodid

Peatükk on kirjutatud toetudes Kuhn jt kirjutatud *caret* paketi dokumentatsioonile (2022) ning Grandini jt artiklile (2020).

Kategooriliste väärtuste imputeerimisel muutub meetodite võrdlus keerulisemaks. Ei saa kasutada üldiselt pidevatele arvulistele tunnustele omaseid erinevusmõõdikuid nagu näiteks ruutjuurt ruutkeskmise veast. Samuti, kuna bakalaureusetöö imputeeritaval kategoorilisel tunnusel on üle kahe väärtuse, on raskendatud üldiste klassifikatsiooni hindavate meetodite

kasutus. Siiski on võimalik hinnata imputeerimismeetodi täpsust (*accuracy*) ning meetodi kiirust. Analüüsiks kasutati *caret* paketi meetodit *confusionMatrix* (eksimismaatriks), mis leiab iga klassi kohta F_1- skoori, saagise ning üle kõigi klasside Coheni kappa kordaja. Pakett kasutab täpsuse ning teiste mõõdikute arvutamiseks *one versus all* lähenemist, kus iga faktortunnuse taseme analüüsis võetakse arvesse kõiki teisi tasemeid.

Täpsus on kõige levinum klassifikaatori headuse hindaja, mis näitab, kui hästi klassifikaator klasse arvestamata väärtuseid hindab, mitmeklassilise kategoorilise tunnuse puhul on täpsus eksimismaatriksi diagonaali summa jagatud eksimismaatriksi summaga. Mõõdik võtab väärtuseid vahemikust [0,1] ning mida ühele lähedasem hinnang on, seda täpsem on klassifikaator. Täpsust on hea kasutada juhtudel, kui ei ole täiesti oluline, mis klassides toimub vaid on vaja üldisemat pilti. Meetodisiseseks võrdlemiseks piisab täpsusest, kuid meetoditevahelise võrdluse puhul tasub arvestada ka teisi mõõdikuid. *ConfusionMatrix* arvutab ka 95% usaldusintervalli kasutades binoomtesti.

Mitme klassiga kategooriliste tunnuste puhul, kus klassid ei ole tasakaalus (lisa 1), on hea kasutada klassifikaatorite võrdluseks Coheni kappa kordajat.

$$K = \frac{c \cdot s - \sum_k^K p_k \cdot t_k}{s^2 - \sum_k^K p_k \cdot t_k}, \text{ kus}$$

$c = \sum_k^K C_{kk}$, õigesti kategoriseeritud väärtuste arv (maatriksi diagonaal);

$s = \sum_i^K \sum_j^K C_{ij}$, kirjete koguarv

$p_k = \sum_i^K C_{ki}$, veeru *k* summa ehk mitu korda puuduvaks väärtuseks klassi *k* hinnati

$t_k = \sum_i^K C_{ik}$, rea *i* summa ehk mitu korda klass *k* tegelikult esineb.

Kordaja eelis ainult täpsuse kasutamise ees on võime kahte klassifikaatorit võrrelda, võttes arvesse ka klasside vahelise ebavõrdsuse (*imbalance*). See annab üldise ülevaate meetodi võimest klassifitseerida ning seejärel imputeerida iga klassi kohta. Coheni kappa kordaja eesmärk on mitte arvestada klassifikaatori juhuslike õigete hinnangutega. Kordaja väärtus on vahemikus [-1,1], kus väärtused alla nulli näitavad, et juhuslik hinnang oleks olnud täpsem kui meetodi hinnang. Kordaja väärtus üks näitab imputeeritud ning tegelike andmete ideaalset kattuvust.

Lisaks Coheni kappale kasutatakse klassifikaatorite võrdlemiseks ka F_1-skoori, mitme klassiga kategoorilise tunnuse puhul kasutatakse Makro F_1-skoori. Skoor näitab esitustäpsuse (*precision*) ning saagise (*recall*) harmoonilist keskmist. Esitustäpsus on tõsipositiivsete (*true positive*) ehk väärtuste, kus klassi väärtust klassifitseeriti õigesti, osakaal kõiki seda klassi ennustanud väärtustest (tõsipositiivsete ning väärpositiivsete summa). Saagis on tõsipositiivsete osakaal üle tõsipositiivsete ning klassi väärnegatiivsete summa, mis näitab, kui tõenäoliselt klassis olev väärtus õigesti klassifitseeritakse.

Klassis *k* arvutatakse esitustäpsust ning saagist:

$$Esitustäpsus_k = \frac{tõsipositiivsed_k}{tõsipositiivsed_k + väärpositiivsed_k}$$

$$Saagis_k = \frac{tõsipositiivsed_k}{tõsipositiivsed_k + väärnegatiivsed_k}$$

Tabelis 1 on näide kolme klassiga eksimismaatriksist ning värviliselt on märgitud klassi esitustäpsuse ning saagise arvutamiseks vajalikud väljad.

Tabel 1. Mitme klassiga kategoorilise tunnuse eksimismaatriks

		Tegelik väärtus		
		30	60	90
Hinnang	30	tõsinegatiivne	väärnegatiivne	tõsinegatiivne
	60	väärpositiivne	tõsipositiivne	väärpositiivne
	90	tõsinegatiivne	väärnegatiivne	tõsinegatiivne

Makro skoori kasutamiseks on vaja igale klassile leida esitustäpsus ning saagis ning võtta nende aritmeetilised keskmised. Esitustäpsuste ning saagiste keskmise pealt arvutatakse Makro F_1- skoor.

$$Makro\ F_1 = 2 \cdot \left(\frac{keskmine\ esitustäpsus \cdot keskmine\ saagis}{keskmine\ esitustäpsus + keskmine\ saagis}\right).$$

Skoori väärtused on vahemikus [0,1], kus null on halvim väärtus, üks parim. Juhul, kui meetod imputeerib vaid kõige sagedasemat klassi õigesti, on F_1- skoor madal (Kuhn jt, 2022; Grandini jt, 2020).

Meetodeid rakendati ning analüüsiti andmetel, kus olid kõik imputeeritavad väärtused teada, imputeerimise jaoks eemaldati igast ravimigrupist juhuslikult ligikaudu 10% ,,annustamise päevi“ väärtustest. Töös edaspidi mainitava andmestiku all peetakse silmas puhastatud andmestikku, milles on puuduvad väärtused ise tektitatud.

4. Meetodite analüüs

Selleks, et valida igasse meetodisse parimad tunnused imputeeritavat väärtust hindama, analüüsiti tunnuseid eraldi. Esialgu kasutati tunnuse imputeerimiseks vaid andmestikus olemasolevaid väljasid („vanus väljakirjutamisel", „sugu") ehk baasi. Kõiki teisi juurde arvutatud tunnuseid analüüsiti koos baasitunnustega, mitte eraldiolevatena. Hinnati diagnooside tunnuseid ehk J44, J45 ning J46, korduvretseptide arvu („samu retsepte päevas"), isikule kokku välja kirjutatud sama ATC-ga retseptide arvu („kokku välja kirjutatud") ning tunnust, mitu sama ATC koodiga retsepti oli isikule varasemalt välja kirjutatud („mitu retsepti enne"). Tunnuste mõju täpsusele hinnati *hot deck*, kNN ja otsustusmetsa algoritmide puhul eraldi.

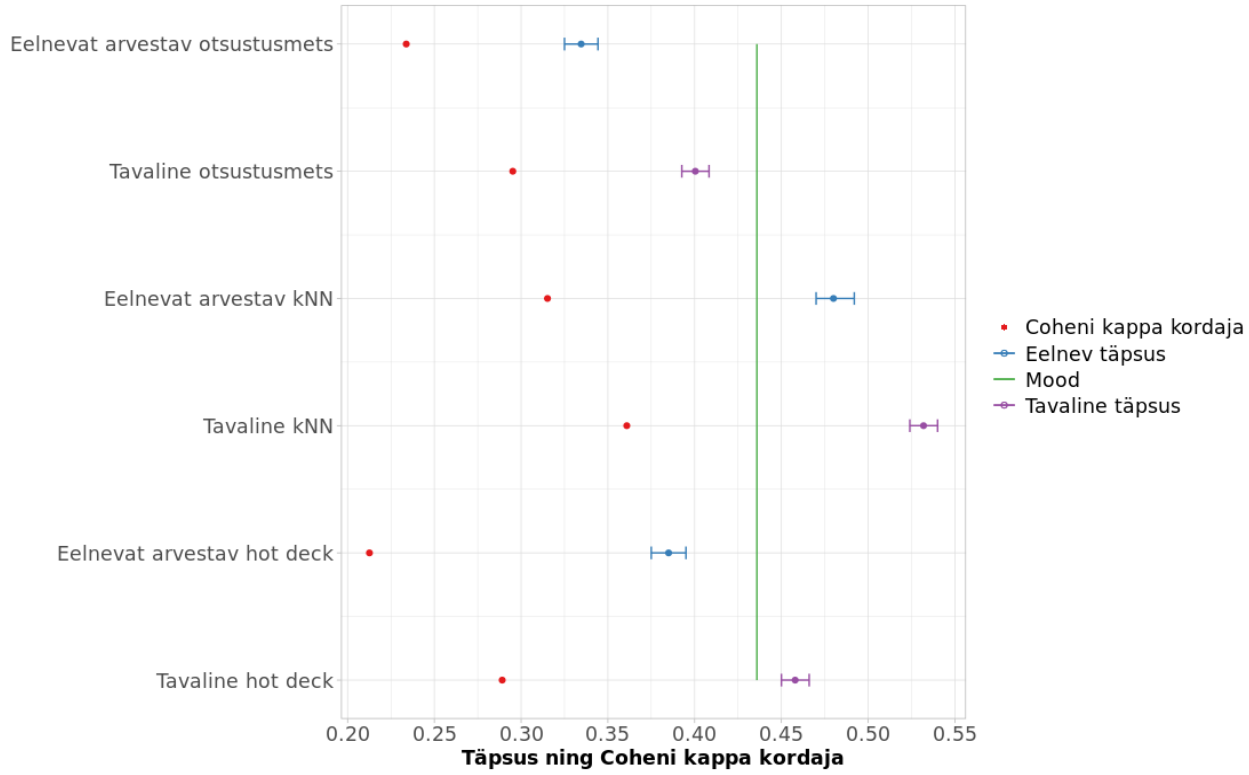
Hot deck meetodit kasutades ilmnes, et järjestamistunnusteks tuleks kindlasti kaasata tunnused „samu retsepte päevas" ning „kokku välja kirjutatud" (lisa 2). Kuna ei saa väita, et teiste tunnustega rakendatud meetodi täpsused erineksid teineteisest oluliselt, prooviti läbi kõikvõimalikud järjestustunnuste kombinatsioonid. Lõpuks ilmnes, et kõikide tunnuste järjestustunnustena lisamine andis täpsuseks 0,458 95% usaldusintervalliga (0,45; 0,466).

KNN meetodi puhul olid samuti olulised tunnused „samu retsepte päevas" ning „kokku välja kirjutatud", kuid kNN puhul erines diagnoositunnustest ning ainult baasitunnustest ka tunnus „mitu retsepti enne" (lisa 2). KNN meetodisse otsustati samuti kaasata kõik tunnused, täpsusega 0,535 ning 95% usaldusintervalliga (0,527; 0,546).

Kuna kNN teeb imputeerimisel valiku mitme võimaliku doonori vahel, ei anna see alati samu tulemusi, seega kõikide headusmõõdikute puhul on tegemist üldise hinnanguga, mis võib vähesel määral varieeruda. Sarnaselt kNN meetodile on ka otsustusmetsa imputeerimismeetod juhuslik, kuid *hot deck* meetod teeb juhusliku valiku vaid siis, kui võimalikke doonoreid on mitu.

Otsustusmetsa imputeerimismeetodi puhul baastunnustele tunnuste „kokku välja kirjutatud" ning „mitu retsepti enne" lisamine muutsid meetodi paremaks. Samas ei olnud „mitu retsepti enne" tunnusel erinevust „samu retsepte päevas" tunnuse täpsusega. Üldiselt oli otsustusmets väheste abitunnuste valikul üsna ebatäpne (lisa 2). Ka otsustusmetsaga imputeerimisel kasutati kõiki tunnuseid, mida kasutades imputeeris meetod täpsusega 0,401 95% usaldusintervalliga (0,394; 0,41).

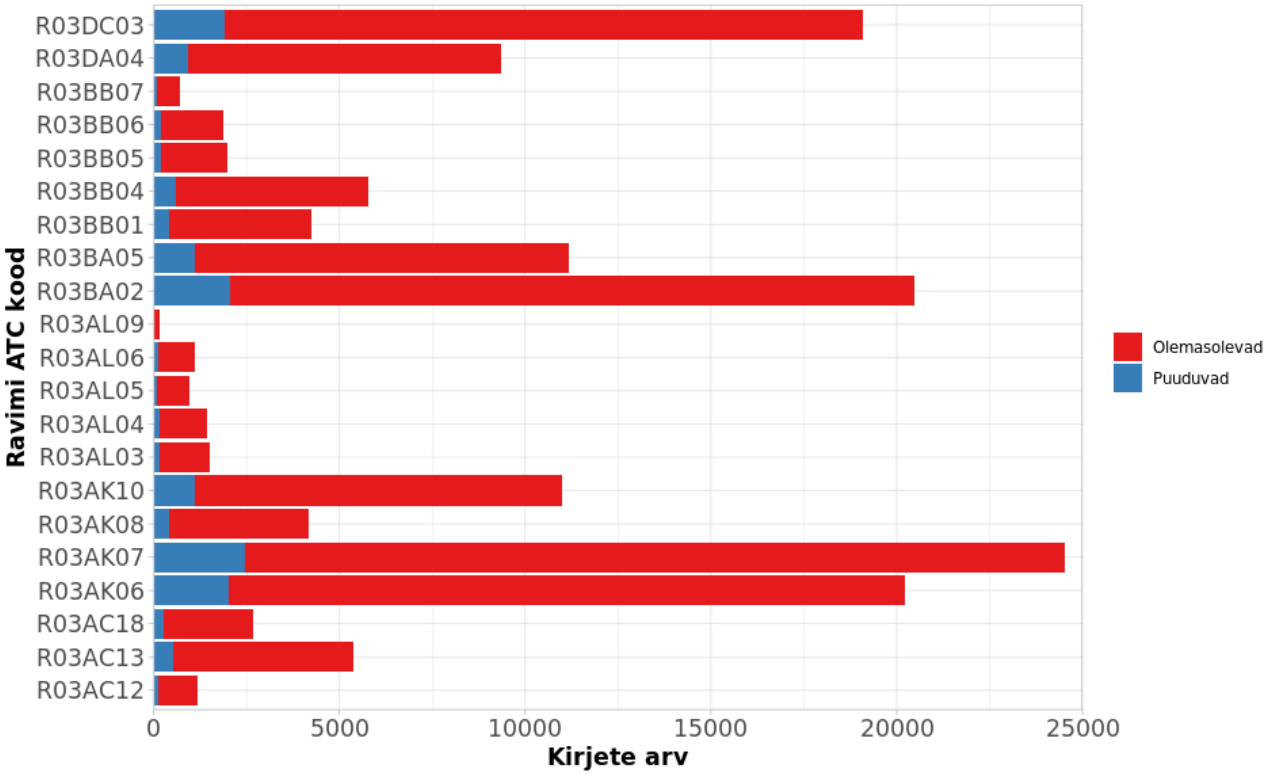
Selleks, et imputeerimismeetodeid parandada, otsustati vaadelda andmeid patsiendi tasemel: kui patsiendile on eelnevalt sama ravimit välja kirjutatud, võib eeldada, et edaspidi kirjutati ravimit sarnasel moel välja. Leiti iga retsepti kohta eelnevalt väljakirjutatud retsepti annustamise päevade arv ning kui eelnevat ravimit kirjutati päevas sama palju välja (ehk mitmekordne retsept), siis arvestati see väärtus puuduoleva annustamise päevade arvu asemele. Korduvretseptide arvu arvestati, kuna ühekordseid retsepte kirjutatakse ravimi võtmise alguses välja, kui patsiendi raviskeem ei ole veel täiesti kindel.



Joonis 3. Täpsuse võrdlus patsiendi raviajalugu arvestades ning mittearvestades

Imputeerimismeetodite täpsust eelnevate retseptidega arvestamine ei parandanud (joonis 3). Rohelise joonega on märgitud iga ATC grupi siseselt moodiga asendades saadud täpsus 0,44, mida kasutatakse analüüsis imputeerimismeetodi täpsuse üldise baashinnanguna. Lillaga on märgitud tavalised meetodid, sinisega eelnevat patsiendi ravimi väljakirjutamise ajalugu arvestavad retseptid. Punase punktina on märgitud Coheni kappa kordaja. Kuigi *hot deck* meetodil näib olevat parem üldine täpsus, oli Coheni kappa kordaja tavalise otsustusmetsa puhul parem, mis viitab otsustusmetsa natukene paremale klassifitseerimisoskusele väiksemates klassides.

Meetodite omavaheliseks võrdluseks analüüsiti iga meetodi puhul iga ravimi tasemel täpsust, ning F_1-skoori. Kui hakata meetodeid vaatama ravimi tasemel, on oluline märkida, et analüüsitavas andmestikus on erinevaid ravimeid erinevas koguses (joonis 4).

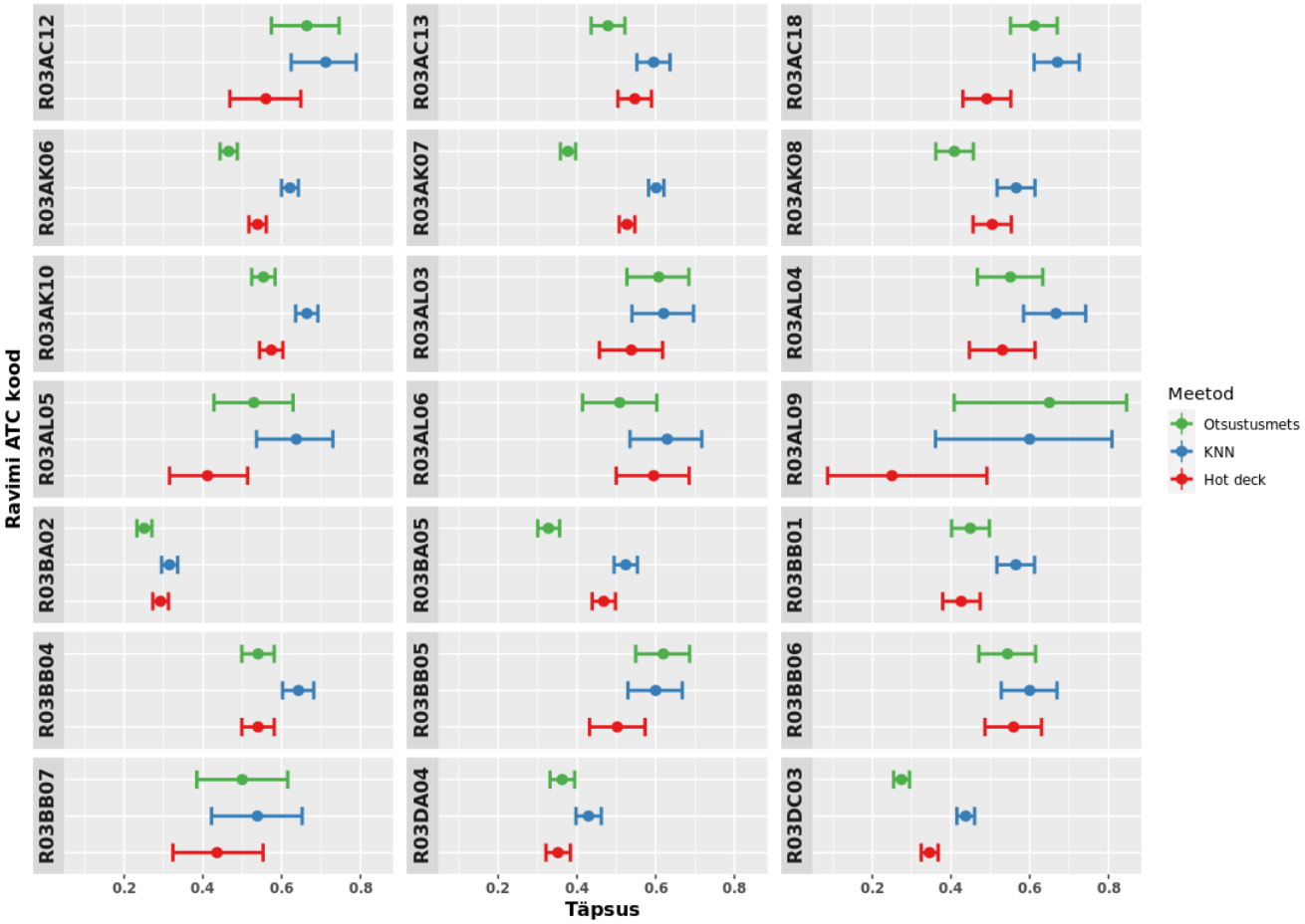


Joonis 4. Imputeeritava andmestiku kirjete arv

Andmestik ei olnud ravimi tasemel tasakaalus, mistõttu võivad väiksemate andmemahtude peal täpsemini imputeerivad meetodid grupiti paremad tunduda kui üle kõikide gruppide vaadates. Kõiki imputeeritud andmeid vaadates oli otsustusmetsa täpsus kõige halvem, kuid oli vaid paar ravimit, mille tasemel sai öelda, et otsustuspuu täpsus oli halvem kui *hot deck* meetodil (joonis 5). Kõik need ravimigrupid olid kirjete poolest mahukad, mistõttu oli neil lõplikus täpsuse arvutamises suurem kaal. KNN meetod oli ka ravimi tasemel teiste meetoditega võrreldes kas parema või sama täpsusega.

Ravimigruppide tasakaalu kõrval tasub vaadata ka klassifitseeritavaid klasse ehk tunnuse „annustamise päevi" erinevaid väärtuseid (lisa 1), mis on üsna ebaühtlaselt jaotunud. Kõige sagedamini esines klass „60" (ligi 42%), mis esines peaaegu sama tihti kui kõik teised klassid kokku. Kui võrrelda kahte ravimigruppi „R03BA02" ja „R03AK06", kus oli ligikaudu sama palju imputeeritavaid andmeid, on näha, et „R03BA02" ravimigrupis olid „annustamise päevi" väärtused ühtlasemalt jaotunud kui „R03AK06" ravimigrupis (lisa 3). Meetodid hindavad

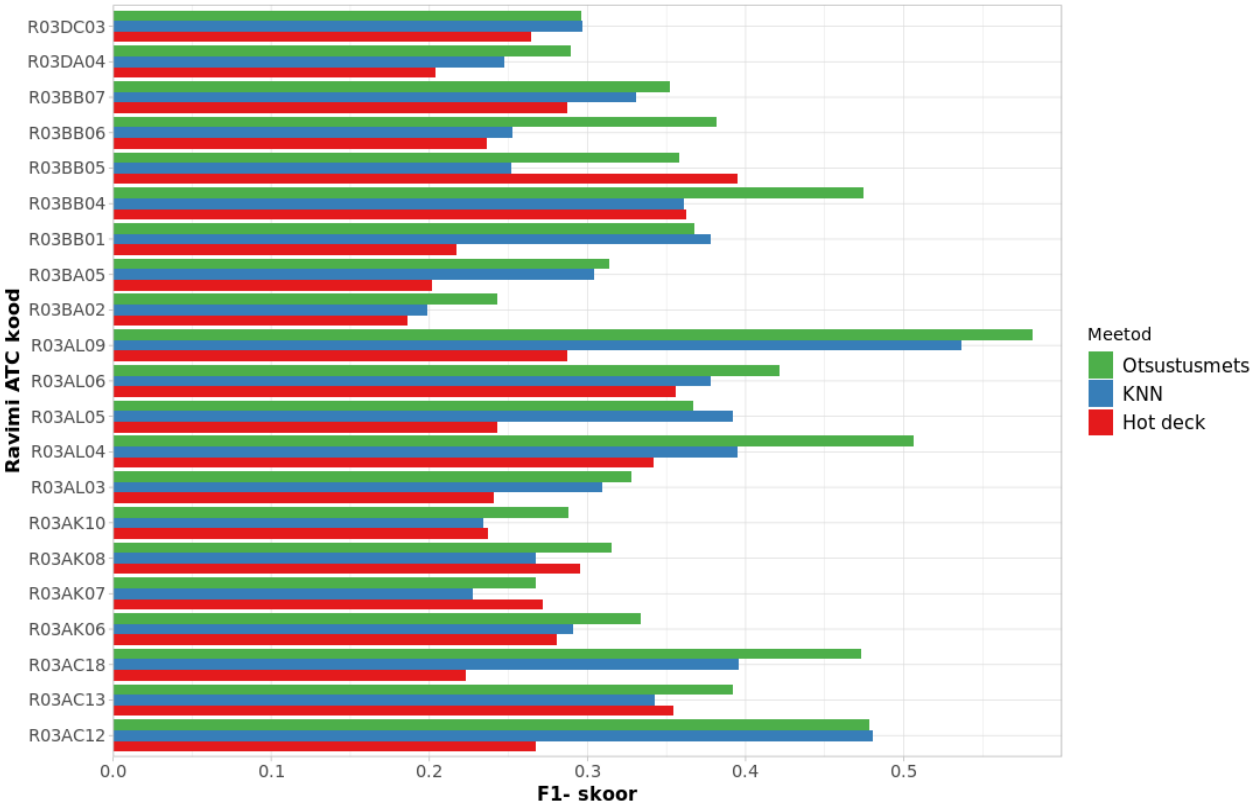
„R03AK06“ gruppi täpsemalt, kuna eelistatakse kõige sagedasemat väärtust. Samas „R03BA02 grupis ei saanud väita, et meetodite täpsused erineksid, mida aga grupis „R03AK06“ oli võimalik järeldada. Sellest võis järeldada, et meetodid imputeerivad sarnaselt ühtlasemalt jaotunud ravimigruppides ning erinevalt ebaühtlaselt jaotunud gruppides.



Joonis 5. Meetodite täpsus ravimigrupiti

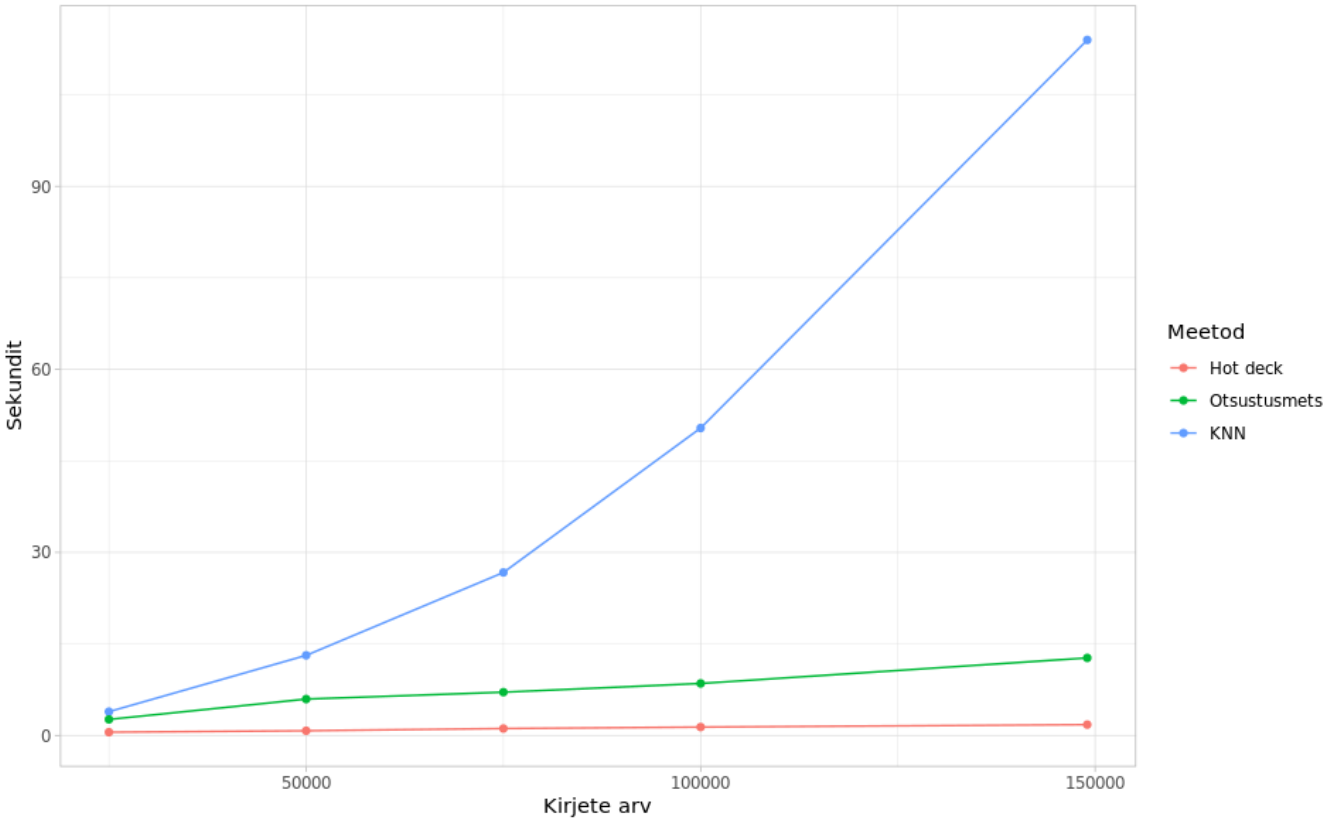
F_1- skoor oli otsustuspuul märgatavalt parem kui *hot deck* meetodil, mis viitab sellele, *et hot deck* meetod imputeeris suures osas kõige sagedasemaid väärtuseid, kuid imputeeris harvem väiksemaid klasse (joonis 6). Ravimigruppide puhul, kus klassi väärtused olid ebaühtlaselt jaotunud, tõstis see küll meetodi täpsust, kuid F_1-skoor langes väiksemate klasside arvelt. Otsustuspuu klassifitseeris tihedamini väiksemaid klasse, mistõttu oli meetodi F_1-skoor kõrgem, kuid üleüldist klasside saagist vaadates ei tähendanud see alati seda, et need hinnangud oleksid korrektsed (lisa 4, lisa 5). KNN meetodi F_1-skoor oli parem kui *hot deckil*, kuid üldiselt halvem kui otsustusmetsal. Ka KNN meetodi puhul põhjustas, nagu *hot deck* meetodilgi

madalamat F_1-skoori, kuid paremat täpsust suuremate klasside tihedamini klassifitseerimine (lisa 6, lisa 7).



Joonis 6. Meetodite F_1- skoor ravimigrupiti

Imputeerimismeetodite rakendamisel on oluline ka silmas pidada meetodite kiirust, eriti, kui andmestik on 3 miljoni kirje suurune. Selleks, et võrrelda imputeerimismeetodite kiiruseid, võrreldi igat meetodit baassisenditega („sugu“ ning „vanus väljakirjutamise hetkel“) erinevate vaatluste arvuga: 25 000, 50 000, 75 000, 100 000 ning ca 150 000. Antud vaatluste arvud peaksid andma üldise ülevaate, milline on iga algoritmi suhe kirjete arvuga. Andmete delikaatsuse tõttu toimus analüüs serveris ning server ei võimaldanud kNN algoritmi jooksutamist, kui kirjeid, mis imputeerimise võimaldamiseks olid omakorda olid jagatud ATC koodide kaupa eraldi andmestikesse, oli üle 150 000. Algoritmide kirjelduste põhjal on võimalik järeldada, et kõikide algoritmide jooksutamise aeg on sõltuv analüüsitavate vaatluste arvust. Jooniselt 7 võib näha, et kõige kiirem meetod oli *hot deck* (150 000 kirje puhul 1.77 sekundit), mille järel oli otsustusmets (150 000 kirjega 12.7 sekundit) ning eelnevatest kordades aeglasem oli KNN meetod, mis kirjete arvu suurenemisega oli 150 000 kirje puhul aeglasem, kui poolteist minutit (114 sekundit).



Joonis 7. Meetodite jooksmisaeg erinevate andmemahtude korral

5. Lõplik hinnang

Imputeerimismeetodi valik oleneb suuresti sellest, mis analüüsi soovitakse imputeeritud andmestikul läbi viia. Mõnel juhul on oluline, et klassifikaator oleks piisavalt tundlik ning ei imputeeriks kõige sagedasemaid väärtuseid. Muul juhul tuleb silmas pidada, et imputeerimismeetod oleks võimalikult täpne.

Hot deck meetod on kõige robustsem, sellega saab kasutada igat tüüpi andmeid, ka nominaalseid tunnuseid. Meetodi puuduseks on väiksemate klasside vähene imputeerimine, kuid eeliseks jällegi tema kiirus. Otsustusmets on vastupidiselt väiksemate klasside sees hea klassifikaator, kuid üldiselt jääb meetod täpsuselt alla moodiga asendamisele. Puhastamata andmestikul ei ole samas võimalik moodi kasutada, kuna puuduvaid andmeid on üle 85%. Seetõttu on otsustusmets siiski parem valik kui moodiga imputeerimine. Otsustusmetsa puhul on oluline, et kõik tunnused oleksid kas arvulised või kategoorilised. Kõige paremaks meetodiks nii ravimigrupiti kui ka üldisel tasemel oli kNN. KNN meetodi rakendamine oli väga mugav, kuid meetodi teeb halvaks tema aeglus. Töö edasiarenduse võimaluseks olekski vaadata kNN imputeerimismeetodit eraldi. KNN meetodi täiustamiseks on võimalik leida parim *k* väärtus ning parim tunnuste kombinatsioon, et meetod oleks kiirem.

Imputeeritava andmestikuga sarnase andmestiku puhul, kus puuduoleval väärtusel on väga palju klasse ning klassid ei ole tasakaalus, ei eelistaks autor ühtegi paketti, arvestades, et juba ca 10% puuduvate andmetega ei olnud meetodite klassifitseerimisvõime väga hea. On võimalik, et suurema hulga puuduvate andmetega oleksid tulemused erinevad. Seepärast tuleks töö täiustamiseks rakendada meetodeid erinevate puuduvate andmete osakaaludega andmestikes ning võimalusel ka tervel esialgsel andmestikul, mida serveris ei olnud võimalik teha. Samuti oleks hea viia läbi sarnane analüüs lühitoimelistele ravimitele.

Paremate tulemuste saamiseks usub autor, et imputeerimismeetodite kasutamise asemel võiks ise mõnda klassifikatsioonimudelit treenida. Ise mudeli treenimise eelis on võimalus treeningandmeid tasakaalustada, mida imputeerimisel üldiselt ei tehta. Peamiseks kasutatud pakettide imputeerimismeetodite puuduseks peabki autor nende üsna piiratud modifitseerimisvõimalusi.

Kokkuvõte

Bakalaureusetöö eesmärk oli leida imputeerimismeetodid, mida KOK ning astmapatsientide andmetel rakendada ning võrrelda. Terviseandmed, milles on tihti puuduvaid väljasid, kasutatakse mitmetes analüüsides, et saada ülevaadet näiteks ravi muutusest ajas ning ravimisoostumusest. Parima võimaliku imputeerimismeetodi leidmine lihtsustab neil andmetel läbiviidavaid uuringuid. Antud analüüsi põhjal on uurijatel lihtsam valida imputeerimiseks vajalikke tunnuseid ning uuringule vastavat imputeerimismeetodit.

Töös imputeeriti tunnust „annustamise päevi“ ehk mitmeks päevaks patsiendile ravimist jätkub. Meetodite võrdluseks kasutati puhastatud andmestikku, kust eemaldati kõik puuduvate väärtustega kirjed. Imputeerimine viidi läbi 10% puuduvatel andmetel, mis genereeriti juhuslikult.

Imputeerimiseks kasutati *hot deck*, kNN ning otsustusmetsa algoritmidel põhinevad meetodeid. Parim klassifitseerimistäpsus oli kNN meetodil (0,535), kuid ravimigrupiti, kui vaadati imputeeritavaid klasse võrdsetena (kasutades F_1- skoori) osutus otsustusmets parimaks valikuks. Seevastu meetodite kiiruse poolest tuleks eelistada *hot deck* meetodit.

Imputeerimise tegi keerulisemaks andmestiku ebaühtlane jaotus nii ravimigrupiti kui ka imputeerimise tunnuse klassiti. Samuti oli keeruline andmestiku puhastus sisestusvigade tõttu.

Üldiselt töö eesmärk täideti: leiti imputeerimismeetodid, mida oli võimalik puhastatud andmetele lihtsalt rakendada. Samuti lisati andmestikku uusi tunnuseid, mis parandasid imputeerimismeetodite täpsust märgatavalt. Konkreetset parimat imputeerimismeetodit eristada ei olnud võimalik, kuna erinevatel meetoditel on erinevad puudujäägid ning tugevused. Parima meetodi valimine nõuab põhjalikumat analüüsi ning oleneb andmetest ja uurimisküsimusest. Üheks võimalikuks töö edasiarenduseks oleks vaadata süvitsi kNN meetodit ning uurida võimalusi seda kiirendada.

Kasutatud allikad

Andridge, R. R., & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Non-response. *Inl Stat Rev, 78*(1), 40-64. doi:10.1111/j.1751-5823.2010.00103.x

Astma andmete uuringutaotlus. (2020).

Dondes, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10). doi:https://doi.org/10.1016/j.jclinepi.2006.01.014

Eesti Kopsuliit. (2022). *KOK*. Kasutamise kuupäev: 09. 05 2022. a., allikas https://www.kopsuliit.ee/haigused/kok/

European Medicines Agency. (2022). *ATC code*. Kasutamise kuupäev: 09. 05 2022. a., allikas https://www.ema.europa.eu/en/glossary/atc-code

Eurostat. (2014). *Memobust Handbook on Methodology of Modern Business Statistics.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://ec.europa.eu/eurostat/cros/system/files/Imputation-04-T-Donor%20Imputation%20v1.0_2.pdf

Fielding, S., Favers, P. M., McDonald, A., McPherson, G., Campbell, M., & RECORD study group. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, 6(57). doi:https://doi.org/10.1186/1477-7525-6-57

Fürnkranz, J. (2010). *Encyclopedia of Machine Learning*. doi:https://doi.org/10.1007/978-0-387-30164-8_204

Global Initiative for Asthma. (2021). Kasutamise kuupäev: 10. 12 2021. a., allikas https://ginasthma.org/about-us/faqs/

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://arxiv.org/pdf/2008.05756.pdf

Haigekassa. (2017). *Kliinilise auditi „Astma käsitlus esmatasandil" kokkuvõte*. Kasutamise kuupäev: 10. 12 2021. a., allikas https://www.haigekassa.ee/sites/default/files/kvaliteet/Kokkuv%C3%B5te.pdf

Käärik, E. (i.a). *Andmeanalüüs 2 loengukonspekt.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://courses.ms.ut.ee/MTMS.01.007/2019_spring/uploads/Main/AA2_Loengukonspekt_2017.pdf

Kiboneka, A. (2020). The evolving burden of asthma and contemporary advances in management: Implications for clinical practice in Southern Africa. *World Journal of Advanced Research and Reviews*, 8(3). doi:10.30574/wjarr.2020.8.3.0315

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *Americal Political Science Review, 95*(1). doi:https://doi.org/10.1017/S0003055401000235

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software, 74*(7). doi:10.18637/jss.v074.i07

Kuhn, M., Wing, J., Weston, S., Keefer, A. W., Chris, Engelhardt, A., . . . Hunt, T. (2022). *Classification and Regression Training.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://cran.r-project.org/web/packages/caret/caret.pdf

Kuus, K. (2021). Ravimisoostumise ennustamine kroonilistel.

Päll, T., & Maiväli, Ü. (i.a). *Puuduvad andmed*. Kasutamise kuupäev: 09. 05 2022. a., allikas Bayesi statistika kasutades R keelt: https://rstats-tartu.github.io/bayesiraamat/puuduvad-andmed.html

Ravijuhendite Nõukoda. (2020). Täiskasvanute astma käsitlus esmatasandil. *RJ-J/3.2-2020*.

Riigi Teataja. (2005). Ravimite väljakirjutamise ja apteekidest väljastamise tingimused ja kord ning retsepti vorm. Kasutamise kuupäev: 09. 05 2022. a., allikas https://www.riigiteataja.ee/akt/106012021015

Ripley, B. (2021). *Classification and Regression Trees.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://cran.r-project.org/web/packages/tree/tree.pdf

Sotsiaalministeerium. (i.a). *Rahvusvaheline haiguste klassifikatsioon*. Kasutamise kuupäev: 09. 05 2022. a., allikas https://rhk.sm.ee/

Stekhoven, D. J. (2022). *Nonparametric Missing Value Imputation using Random Forest.* Kasutamise kuupäev: 04. 05 2022. a., allikas https://cran.r-project.org/web/packages/missForest/missForest.pdf

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - nonparametric missing value imputation for. *BioInformatics, 28*(1), 112-118. doi:https://doi.org/10.1093/bioinformatics/btr597

Sterne, J. A., Carlin, J. B., Royston, P., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ, 338*. doi:https://doi.org/10.1136/bmj.b2393

Templ, M., Kowarik, A., Alfons, A., de Cillia, G., Prantner, B., & Rannetbauer, W. (2021). *Visualization and Imputation of Missing Values.* Kasutamise kuupäev: 09. 05 2022. a., allikas https://cran.r-project.org/web/packages/VIM/VIM.pdf

van Buuren, S. (i.a). Flexible imputation of Missing Data. Kasutamise kuupäev: 04. 05 2022. a., allikas https://stefvanbuuren.name/fimd/

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine, 4(1)*, 9. doi:10.3978/j.issn.2305-5839.2015.12.38

Lisad

Lisa 1. Tunnuse ,,annustamise päevi" klasside jaotus enne andmestiku puhastamist ning ise puuduvate väärtuste tekitamist

Klass	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**	∑
Kirjeid	3271	10949	782	640	3061	1647	16029	6269	32289	2149	3390	61929	3262	3293	148960
Sagedus	~0,022	~0,074	~0,005	~0,004	~0,021	~0,011	~0,108	~0,042	~0,217	~0,014	~0,023	~0,416	~0,022	~0,022	~1

Lisa 2. Lisatud tunnustega meetodi täpsus ning 95% usaldusintervall

Lisatud tunnused / *Meetod*	**Baas**	**Baas+ diagnoosid**	**Baas+ ,,Samu retsepte päevas"**	**Baas+ „Kokku välja kirjutatud"**	**Baas + „Mitu retsepti enne"**
Hot deck	0,344 (0,335;0,351)	0,348 (0,34;0,356)	0,421 (0,413;0,429)	0,408 (0,399;0,416)	0,352 (0,345;0,36)
KNN	0,3242 (0,317;0,332)	0,3397 (0,332; 0,347)	0,4526 (0,445;0,461)	0,4426 (0,435;0,451)	0,3932 (0,385;0,401)
Otsustusmets	0,0306 (0,028;0,034)	0,0489 (0,046;0,053)	0,1594 (0,154;0,165)	0,2908 (0,284;0,298)	0,158 (0,152;0,164)

Lisa 3. Imputeeritavate väärtuste jaotus puhastatud andmetel ravimigrupiti

Ravimi-grupp **Klass**	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**	***NA***	∑
R03AC12	4	28	100	120	1	31	147	0	109	0	0	715	0	21	125	1401
R03AC13	102	269	1	44	24	60	623	1	704	3	0	2848	18	119	543	5359
R03AC18	10	135	0	1	2	21	332	0	884	0	0	1025	0	19	273	2702
R03AK06	95	770	10	75	71	297	2790	7	3733	14	16	9902	25	385	2028	20218
R03AK07	175	922	7	236	85	306	2898	10	4011	10	10	12826	72	505	2464	24537
R03AK08	38	189	0	16	10	63	319	0	734	2	0	2025	1	352	426	4175
R03AK10	74	621	0	27	13	93	1038	6	4484	4	4	3411	7	93	1105	10980
R03AL03	12	110	0	1	0	6	121	0	536	0	0	549	0	15	158	1508
R03AL04	2	68	0	0	0	17	177	0	395	0	0	620	0	13	147	1439
R03AL05	0	59	0	3	0	9	81	0	290	0	0	394	0	23	102	961
R03AL06	4	55	0	9	0	8	115	0	301	0	0	474	0	14	116	1096
R03AL09	0	5	0	0	0	5	30	0	36	0	0	63	0	0	20	159
R03BA02	1673	3443	158	31	1212	109	1756	58	3120	531	4	3851	2188	293	2056	20483
R03BA05	108	771	3	47	218	63	1069	25	3307	23	2	4188	85	149	1126	11184
R03BB01	254	313	18	4	63	50	333	2	729	45	0	1779	153	72	432	4247
R03BB04	20	288	7	11	5	41	775	0	1376	2	0	2604	6	45	585	5765
R03BB05	8	88	0	1	1	7	229	0	647	0	1	776	0	14	205	1977

R03BB06	9	83	0	2	1	13	223	0	460	0	0	864	1	20	195	1871
R03BB07	0	36	0	3	0	3	121	0	257	0	0	210	0	10	78	718
R03DA04	209	720	485	16	310	164	667	1	1007	1265	0	3145	226	188	941	9344
R03DC03	116	871	8	36	728	98	573	5526	1972	26	3008	3459	145	602	1916	19084
∑	2913	9854	897	803	2758	1644	14437	5664	29122	1975	3101	55788	2934	3042	15041	149973

Lisa 4. Saagis klassiti

Klass Meetod	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**
Hot deck	0.196	0.25	0,067	0,052	0,160	0,046	0,313	0,47	0,495	0,188	0,26	0,592	0,171	0,175
KNN	0.296	0.3	0.06	0.069	0.18	0.072	0.378	0.501	0.551	0.244	0.294	0.61	0.208	0.256
Otsustus-mets	0.201	0.221	0.063	0.052	0.115	0.06	0.336	0.295	0.596	0.181	0.277	0.796	0.123	0.12

Lisa 5. Otsustuspuu eksimismaatriks

Hinnang / Tegelik väärtus	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**	∑
≤ 5	100	25	8	0	40	11	22	28	34	8	3	21	42	16	358
10	48	342	38	8	109	24	194	37	58	30	4	77	111	25	1115
100	3	1	27	1	11	11	0	1	9	2	0	10	9	0	185
120	0	3	1	14	9	5	12	1	18	0	0	9	2	0	194
14	5	5	14	2	136	13	2	24	41	8	3	1	43	20	331
180	0	2	2	0	9	59	2	2	57	3	0	12	11	15	354
20	46	325	15	8	51	13	846	15	30	17	2	193	44	7	1632
28	15	1	0	6	149	37	2	152	19	13	30	11	63	135	661
30	60	78	99	35	312	253	70	77	1676	30	8	142	124	273	3267
50	8	8	26	1	19	26	13	3	20	52	0	19	16	13	274
56	13	38	0	4	18	28	42	41	5	3	49	15	28	61	401
60	169	691	76	72	215	290	1284	62	724	63	17	2277	105	156	6261
7	30	18	19	1	69	12	4	35	29	8	1	5	86	18	342
90	1	3	1	2	20	15	0	9	63	0	4	8	8	207	431
∑	498	1550	426	274	1181	977	2513	515	2813	287	177	2860	699	1036	15806

Lisa 6. *Hot deck* eksimismaatriks

Hinnag / Tegelik väärtus	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**	∑
≤ 5	72	50	5	0	7	2	38	5	59	8	4	68	31	9	358
10	48	274	8	8	40	4	125	9	132	15	20	351	57	14	1115
100	5	12	12	0	1	3	5	0	15	12	0	18	2	0	185
120	2	4	0	10	2	0	6	0	18	2	0	26	3	1	194
14	19	48	2	1	53	3	13	35	56	10	5	40	25	7	331
180	2	4	4	2	0	16	6	7	65	3	0	54	7	4	354
20	36	150	5	7	21	4	518	7	102	12	8	707	21	14	1632
28	6	13	0	0	42	4	5	305	62	1	102	67	8	18	661
30	44	104	9	9	67	48	103	74	1662	37	33	886	54	107	3267
50	8	17	9	1	8	6	16	1	34	51	0	59	10	4	274
56	2	15	0	0	6	3	17	85	40	0	108	61	1	7	401
60	85	317	20	29	42	66	743	60	921	64	66	3677	42	69	6261
7	33	67	3	1	23	2	34	9	53	5	0	45	56	4	342
90	5	12	2	1	4	7	4	24	108	2	14	82	2	74	431
∑	367	1097	179	189	330	348	1653	649	3357	272	416	6201	326	422	15806

Lisa 7. KNN meetodi eksimismaatriks

Hinnang / Tegelik väärtus	**≤ 5**	**10**	**100**	**120**	**14**	**180**	**20**	**28**	**30**	**50**	**56**	**60**	**7**	**90**	∑
≤ 5	59	48	0	0	8	1	20	3	73	5	8	111	21	1	358
10	27	240	3	1	30	4	136	12	137	19	14	426	50	6	1115
100	0	4	8	0	0	0	3	1	27	10	0	30	2	0	185
120	0	3	0	9	0	0	3	0	22	0	1	36	0	0	194
14	6	42	1	0	30	0	16	49	78	7	6	53	22	7	331
180	1	3	2	0	1	17	2	5	61	1	1	77	2	1	354
20	20	106	2	0	12	0	452	3	56	11	12	920	17	1	1632
28	0	3	0	0	19	1	1	410	39	0	75	55	1	29	661
30	21	83	3	0	19	9	48	85	2002	17	21	830	25	74	3267
50	6	11	8	0	1	1	9	3	38	53	0	81	4	9	274
56	1	7	0	0	1	0	8	114	26	0	107	67	0	14	401
60	31	162	4	1	13	17	453	70	846	34	47	4478	20	25	6261
7	25	75	0	0	16	3	24	9	71	7	3	55	46	1	342
90	2	7	0	0	3	2	1	26	127	3	13	64	4	89	431
∑	199	804	131	131	167	235	1196	818	3633	217	364	7343	221	347	15806

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Agnes Annilo,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „ Krooniliste astmahaigete pikatoimeliste ravimiandmete imputeerimine”, mille juhendaja on Raivo Kolde, 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.

Agnes Annilo
10.05.2022