Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kadi Kilgi

Hospitaliseerimise riski prognoosimine krooniliste haigustega patsientidel

Kindlustus- ja finantsmatemaatika

Magistritöö (30 EAP)

Juhendajad: Mark Gimbutas, MSc
Prof. Krista Fischer, PhD

Tartu 2022

HOSPITALISEERIMISE RISKI PROGNOOSIMINE KROONILISTE HAIGUSTEGA PATSIENTIDEL

Magistritöö

Kadi Kilgi

Lühikokkuvõte

Käesoleva magistritöö eesmärk on eelmise aasta riskipatsientide raviarvete andmeid kasutades prognoosida järgmisel aastal välditavat hospitaliseerimist vajavad patsiendid. Riskipatsiendiks loetakse krooniliselt haiget inimest, kellel on suurenenud risk tervise halvenemisele. Lisaks on soov mudeliga hinnata patsientidele hospitaliseerimise riskiskoor, mille alusel patsiendid järjestada. Töö teoreetilises osas tutvustatakse masinõppe metoodikat ning kirjeldatakse töös kasutatavaid klassifitseerimismeetodeid. Lisaks tehakse ülevaade tasakaalustamata andmete probleemist ning võimalikest lahendustest. Seejärel tehakse ülevaade riskipatsientide definitsioonist ning kirjeldatakse valimisse sattunud patsiente. Töö praktilises osas katsetatakse erinevaid klassifitseerimismeetodeid ning võrreldakse erinevaid lähenemisi hospitaliseerimiste prognoosimisel. Töö tulemusena valitakse parim meetod ning katsetatakse valitud mudelit uute andmete korral.

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

Märksõnad: Tehisõpe, üldistatud lineaarsed mudelid, otsustusmets, närvivõrk, prognostika.

PREDICTING HOSPITALIZATION RISK AMONG PATIENTS WITH CHRONIC DISEASES

Master thesis

Kadi Kilgi

Abstract

The aim of this master's thesis is to use data of medical claims from last year about high-risk patients and find a model that would predict avoidable hospital admissions next year. High-risk patients are patients with chronic diseases, whose health might worsen. In addtion, the model should give a patient a risk score for hospitalization, which could be used to order the high-risk patients. In theoretical part, the concept of machine learning is introduced. Also, the classifiers that are used for modelling are presented. Furthermore, the problem of imbalanced data is pointed out with possible solutions. Next, the definition of high-risk patient is described. In addition, a general description of a patient in the dataset is given. In the practical part, different classifiers are compared for modelling hospital admissions and the best model is selected. At the end, a final best model is chosen and it's performance is evaluated in case of new data.

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

Key Words: Machine learning, generalized linear models, random forest, neural networks, prediction.

Sisukord

Sissejuhatus

Käesoleva töö raames on vaatluse all vähemalt ühe kroonilise haiguse diagnoosiga (hüpertensioon, hüperlipideemia, diabeet) patsiendid. Neil on tihti veel teisi kaasuvaid haiguseid ning nad kasutavad mitmeid erinevaid ravimeid. Lisaks vajavad nad haiguste ägenemisel erinevaid tervishoiuteenuseid. Oma terviseseisundi tõttu on neil suurenenud risk tervise halvenemisele. Kirjeldatud patsiente nimetatakse riskipatsientideks. (Šteinmiller, 2021) Uuringuga (Eesti Haigekassa ja Maailmapanga Grupp, 2015) on kindlaks tehtud, et perearsti sekkumisega on võimalik riskipatsientide kroonilise haiguse halvenemisest tingitud välditavaid hospitaliseerimisi ära hoida.

Magistritöö eesmärk on eelmise aasta riskipatsientide raviarvete andmeid kasutades prognoosida järgmisel aastal välditavat hospitaliseerimist vajavad patsiendid. Leitava mudeli eesmärk on patsiendid klassifitseerida kõrge ja madala hospitaliseerimise riskiga patsientideks. Lisaks klassifitseerimisele on soov patsiendid järjestada mõne riskiskoori alusel, näiteks hospitaliseerimise tõenäosuse järgi. Riskiskoori kasutades saab riskipatsiendid järjestada ning jälgimise alla võtta just kõrgema riskiga patsiendid, kellel on risk järgmisel aastal välditavaks hospitaliseerimiseks.

Töö koosneb kolmest peatükist. Esimeses peatükis antakse ülevaade masinõppe metoodikast ning kirjeldatakse klassifitseerimismeetodeid, mida töös kasutatakse. Lisaks tehakse ülevaade tasakaalustamata andmete probleemist ning võimalikest lahendustest. Teises peatükis kirjeldatakse riskipatsiendi mõistet ning analüüsiks kasutatavaid andmeid. Lisaks tehakse ülevaade andmete töötlusest ning valimisse sattunud riskipatsientidest. Viimases peatükis kirjeldatakse praktilise osa tulemusi. Töö praktilises osas katsetatakse erinevaid klassifitseerimismeetodeid ning võrreldakse erinevaid lähenemisi välditavate hospitaliseerimiste prognoosimisel. Lõpuks valitakse vaadeldud mudelitest välja parim mudel. Seejärel leitakse parima mudeli prognoosid uute andmete korral ning hinnatakse nende täpsust.

Lõputöö on vormistatud tekstitöötlusprogrammiga LaTeX. Praktilises osas kasutatakse rakendustarkvara *R* (R Core Team, 2022b) ja *Python* (Python Software Foundation, 2001-2022).

1 Metoodika

Järgnevas peatükis tutvustatakse analüüsiks vajalikke mõisteid ja meetodeid. Peatükis kasutatud eestikeelsed vasted põhinevad allikatest Tiit ja Tooding (2019) ning Data Science Estonia (2022).

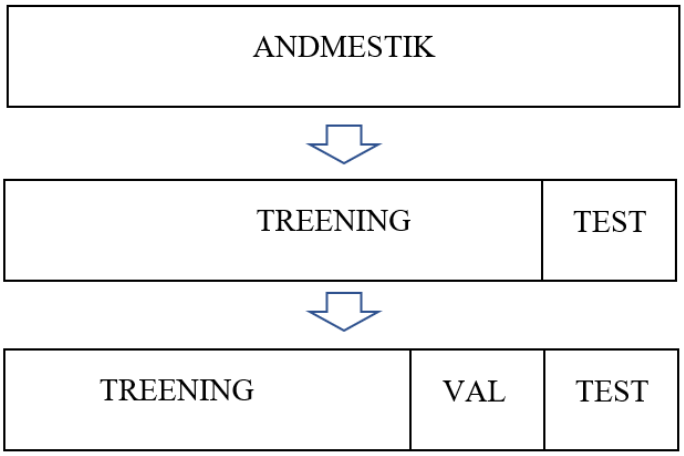
1.1 Masinõpe

Alapeatükk põhineb raamatutel Goodfellow, Bengio ja Courville (2016, ptk 5) ning James *et al.* (2021, lk 15–33, 197–200), kui ei ole märgitud teisiti.

Masinõppe (*machine learning*) meetodid on järjest enam populaarsust koguvad meetodid, mille edu tuleneb andmemahtude kiirest kasvust. Masinõppe eesmärk on andmete pealt õppida ja lahendada soovitud probleem. Ülesande lahendamiseks kasutatakse andmestikku $\mathbf{X} \in \mathbb{R}^{n \times p}$, mis koosneb vaatlustest (*observations, examples*).

Masinõppe algoritmid jagunevad kaheks: juhendatud õpe (*supervised learning*) ja juhendamata õpe (*unsupervised learning*). Erinevus seisneb selles, et juhendatud õppe puhul on lisaks sisendandmetele teada ka uuritava tunnuse väärtused ehk mudeli õpetamisel on ka tulemus teada. Juhendamata õppe puhul ei ole väljund teada, vaid algoritmide abil otsitakse andmestikest seoseid ja kasulikke omadusi. Antud töö käigus keskendutakse juhendatud õppe meetoditele.

Juhendatud õppe meetodeid kasutatakse uutel andmetel prognoosi leidmiseks. Seega mudeli treenimisel on oluline, et hinnatav mudel teeks vähe vigu nii treeningandmetel kui ka uutel andmetel. Selleks, et hinnata mudeli võimekust (*model performance*) uute andmete korral, jagatakse töödeldud andmestik kaheks: treening- ja testandmeteks (vt joonis 1). Treeningandmestikku kasutatakse mudeli parameetrite hindamiseks ning testandmestikku mudeli võimekuse hindamiseks. Oluline on, et testandmestikuga ei oleks mudel treeningprotsessi käigus kordagi kokku puutunud.



Joonis 1: Andmete jaotumine treening-, valideerimis- ja testandmeteks

Andmestiku jagamisel kaheks peab leidma tasakaalu, et mudeli treenimiseks jääks piisavalt andmeid ning et testandmestik ei oleks samal ajal liiga väike.

Masinõppe meetoditel kasutatakse lisaks mudeli parameetritele ka hüperparameetreid (*hyperparameters*), mis määratakse mudeli treeningprotsessist väljaspool. Hüperparameetrite valik mõjutab mudeli sobituvust andmetele, seega on vajalik nende seadistamine. Oluline on, et hüperparameetrite valikul kasutatakse samuti ainult treeningandmestikku, sest see on osa mudeli hindamisest. Hüperparameetrite valimiseks jaotatakse olemasolev treeningandmestik kaheks: treening- ja valideerimisandmestikuks (vt joonis 1). Valideerimisandmestikku kasutatakse hüperparameetrite valimiseks.

Andmestiku jagamiseks saab kasutada rakendustarkvara *R* paketti *rsample* funktsiooni *initial_split*. Antud funktsioonile tuleb ette anda osakaal andmestikust, mida soovitakse kasutada mudeldamiseks. Vaikeväärtusena on treeningandmestiku osakaaluks 0,75. Selleks, et uuritava tunnuse jaotus oleks sama nagu algses andmestikus, saab määrata tunnuse, mille järgi moodustatakse kihtvalim (*stratified sampling*). Kirjeldatud funktsiooni saab kasutada nii test- kui ka valideerimisandmestiku moodustamiseks. (Silge *et al.*, 2021)

1.2 Klassifitseerimine

Järgnev alapeatükk põhineb raamatutel James *et al.* (2021, lk 16–21, 129–133) ning Goodfellow, Bengio ja Courville (2016, lk 97), kui ei ole märgitud teisiti.

Uuritav tunnus võib olla nii kvantitatiivne kui ka kvalitatiivne. Antud töö käigus keskendutakse kvalitatiivsele tunnusele ning uuritakse klassifitseerimismeetodeid (*classification*). Kvalitatiivne tunnus võib olla oma olemuselt nominaalne, binaarne või pidev tunnus, mis on jaotatud kategooriateks. Kvalitatiivne tunnus on näiteks:

- haridustase (põhi-, kesk- või kõrgharidus),
- haigestunud vähki või mitte,
- alla keskmist palka saavad inimesed ja üle keskmist palka saavad inimesed.

Klassifitseerimisülesande korral treenitakse klassifitseerija (*classifier*), mis jagab vaatlused sobivasse klassi k. Klassifitseerija treenimise käigus soovitakse hinnata funktsioon $f : \mathbb{R}^n \to \{1, 2, ..., K\}$, kus K tähistab klasside koguarvu. Funktsiooni f kirjeldamiseks kasutatakse andmestikku, mis koosneb paaridest (X_1, Y_1), (X_2, Y_2),...,(X_n, Y_n), kus $X_i = (X_{i1}, X_{i2}, ..., X_{ip}) \in \mathbb{R}^p$, ning leitakse hinnang $\hat{f}$ nii, et

$$Y_i \approx \hat{f}(X_i).$$

Klassifitseerijaga leitud tulemus võib olla nii konkreetse klassi prognoos kui ka tõenäosus kuuluda mingisse klassi. Antud magistritöö raames on uuritav tunnus binaarne

$$Y_i = \begin{cases} 1, & \text{kui patsient hospitaliseeritakse,} \\ 0, & \text{kui patsiendi hospitaliseerimist ei toimu.} \end{cases}$$

1.3 Üldistatud lineaarsed mudelid

Järgnev alapeatükk põhineb raamatutel Jong ja Heller (2008, lk 20–22, 35–40, 64–70, 97–99) ja James *et al.* (2021, lk 133–139, 170, 237–242).

Üldistatud lineaarsete mudelite korral eeldatakse, et uuritava tunnuse Y_i jaotus kuulub eksponentsiaalsete jaotuste perre ning, et uuritava tunnuse keskväärtus μ_i on seotud kirjeldavate tunnustega mingi teisenduse kaudu. Uuritava tunnuse Y_i jaotus kuulub eksponentsiaalsete jaotuste perre, kui tema tõenäosusfunktsioon (diskreetsetel jaotustel) või tihedusfunktsioon (pidevatel jaotustel) avaldub kujul

$$f(Y_i) = c(Y_i, \phi_i) \cdot \exp \frac{Y_i\theta_i - a(\theta_i)}{\phi_i},$$

kus θ_i on kanooniline parameeter ja ϕ_i tähistab hajuvuse parameetrit. Funktsioonid $a(\theta_i)$ ja $c(Y_i, \phi_i)$ määravad ära, mis jaotusega on tegu (näiteks binoomjaotus). Mudeldamisel eeldatakse, et uuritava tunnuse keskväärtus μ_i on lineaarselt seotud kirjeldavate tunnustega seosefunktsiooni (*link function*) kaudu

$$g(\mu_i) = \beta_0 + \beta_1 \cdot X_{i1} + ... + \beta_p \cdot X_{ip}.$$

Kui uuritav tunnus Y_i on binaarne, $Y_i \in \{0, 1\}$, siis $Y_i \sim B(1, p_i)$, kus p_i tähistab tõenäosust, et $Y_i = 1$. Bernoulli jaotuse korral

$$E(Y_i) \equiv \mu_i = p_i$$
$$D(Y_i) = p_i(1 - p_i).$$

Bernoulli jaotus kuulub eksponentsiaalsesse jaotuste perre ning tema tõenäosusfunktsioon avaldub kujul $f(Y_i) = p_i^{Y_i}(1 - p_i)^{1-Y_i}$.

Logistilise regressiooni korral kasutatakse seosefunktsioonina logitfunktsiooni

$$g(p_i) = \ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \cdot X_{i1} + ... + \beta_p \cdot X_{ip}.$$

Logitfunktsioon kindlustab, et hinnatav tõenäosus jääb vahemikku $(0, 1)$. Logistilise regressiooni korral pakub huvi tõenäosuse hindamine

$$E(Y_i) = p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip})}.$$

Täiend-log-log seosefunktsioon avaldub kujul

$$g(p_i) = \ln\{-\ln(1-p_i)\}$$

ning

$$p_i = 1 - \exp(-\exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip})).$$

Parameetrite $\beta = (\beta_0, \beta_1, ..., \beta_p)$ hindamiseks kasutatakse suurima tõepära meetodit. Suurima tõepära meetodi korral maksimeeritakse log-tõepärafunktsioon

$$l(\beta) = \sum_{i=1}^{n} \ln f(Y_i; \beta, \phi_i) = \sum_{i=1}^{n} \left\{ \ln c(Y_i, \phi_i) + \frac{Y_i\theta_i - a(\theta_i)}{\phi_i} \right\}.$$

Rakendustarkvara *R* korral kasutatakse üldistatud lineaarsete mudelite hindamiseks paketti *stats* ning funktsiooni *glm*. Mudeli parameetrite leidmiseks kasutatakse iteratiivset kaalutud vähimruutude meetodit (*iteratively weighted least squares*). (R Core Team, 2022a)

Tunnuste arvu vähendamiseks on võimalik lisada hinnatavatele parameetritele kitsendused, mis kahandavad parameetrite $\beta_1, \beta_2, ..., \beta_p$ hinnangute väärtused nulli lähedale ning mõned väärtused ka nulli. Seega minimeeritakse negatiivset log-

tõepärafunktsiooni muudetud kujul:

$$\min_{\beta_0,\beta_1,...,\beta_p} \left\{ -l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

kus parameeter λ valitakse valideerimisandmestikku kasutades. Kui $\lambda = 0$, siis parameetrite hinnangute kahandamist ei toimu. Kirjeldatud meetodit nimetatakse **lassoregressiooniks**. Antud töös kasutatakse lassoregressiooniks rakendustarkvara *R* paketti *tidymodels* (Kuhn ja Wickham, 2020) ning funktsiooni *glmnet* (Kuhn ja Vaughan, 2022).

1.4 Otsustuspuu

Järgnev alapeatükk põhineb raamatutel James *et al.* (2021, lk 327–338) ning Hastie, Tibshirani ja Friedman (2009, lk 305–311), kui ei ole märgitud teisiti.

Otsustuspuu ehitamiseks kasutatakse olemasolevaid kirjeldavaid tunnuseid $X_j = (X_{1j}, X_{2j}, ..., X_{nj})^T \in \mathbb{R}^n$, ning tekitatakse nende põhjal J mittekattuvat piirkonda $R_1, R_2, ..., R_J \subset \mathbb{R}^p$. Piirkondi R_m $(m = 1, 2, ..., J)$ nimetatakse lehtedeks (*leaves*, *terminal nodes*). Klassifitseerimise korral leitakse iga piirkonna prognoos vastavalt piirkonda kuuluvatele vaatlustele. Piirkonna prognoosiks saab klass $k = 1, 2, ..., K$, mida esineb piirkonnas kõige rohkem.

Klassifitseerimispuu korral kasutatakse piirkondade leidmisel klassifitseerimisviga E, Gini indeksit G või entroopiat D:

$$E = 1 - \max_k \hat{p}_{mk},$$
$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$
$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk},,$$

kus $\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} \mathrm{I}(Y_i = k)$ tähistab k-nda klassi osakaalu piirkonnas m ning n_m tähistab vaatluste arvu piirkonnas m. Soov on leida piirkonnad nii, et valitud hinnang (E, G või D) oleks minimaalne. Gini indeks ja entroopia on väiksed, kui piirkond m sisaldab rohkelt vaatlusi ühest klassist ehk $\hat{p}_{mk}$ on lähedal nullile või ühele.

Kui tegemist on binaarse klassifitseerimisülesandega, kus $Y_i \in \{0, 1\}$ ning $p_{m1} = P(Y_i = 1 | X_i \in R_m)$, siis saavad Gini indeks ja entroopia kuju:

$$\begin{aligned} G &= \hat{p}_{m0}(1 - \hat{p}_{m0}) + \hat{p}_{m1}(1 - \hat{p}_{m1}) = \\ &= (1 - \hat{p}_{m1})\hat{p}_{m1} + \hat{p}_{m1}(1 - \hat{p}_{m1}) = \\ &= 2\hat{p}_{m1}(1 - \hat{p}_{m1}) \\ D &= -\hat{p}_{m0} \log \hat{p}_{m0} - \hat{p}_{m1} \log \hat{p}_{m1} = \\ &= -((1 - \hat{p}_{m1}) \log (1 - \hat{p}_{m1}) + \hat{p}_{m1} \log \hat{p}_{m1}). \end{aligned}$$

Puu defineerimine algab juurtipust (*root node*). Vaja on leida tunnus X_j ja lõikepunkt s nii, et valitud hinnang (E,G või D) oleks minimaalne. Selliselt tegutsedes saadakse kaks piirkonda:

$$R_1(j, s) = \{X | X_j < s\} \text{ ja } R_2(j, s) = \{X | X_j \geq s\}.$$

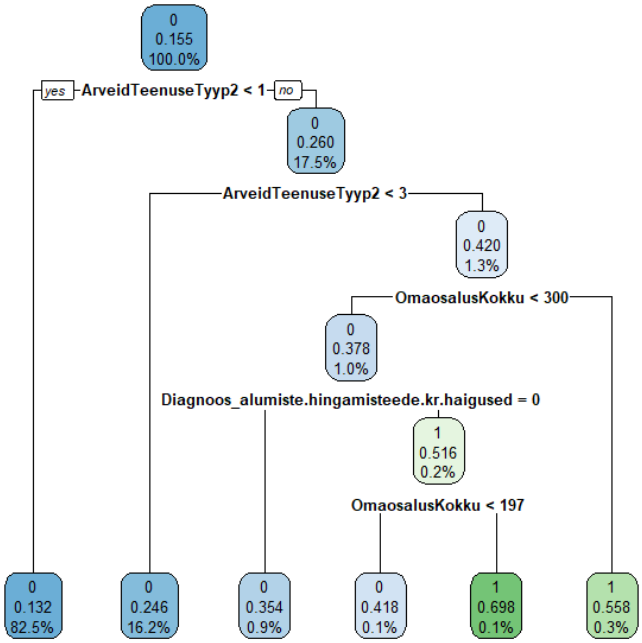
Seejärel leitakse mõlemas piirkonnas kõige sobilikumad tunnused vahetipu jaoks koos lõikekohtadega. Nii jätkatakse, kuni lehed sisaldavad eelnevalt fikseeritud minimaalse arvu vaatlusi.

Jõudes lehtedeni, mis sisaldavad minimaalse arvu vaatlusi, jõutakse suurima puuni T_0. Sellise otsustuspuu puhul on oht treeningandmetele ülesobitamisele (*overfitting*), mis teeb mudeli kasutuks uute andmete korral. Selleks, et leida parim puu suurus, kasutatakse puu pügamist (*prune a tree*), mille käigus leitakse alampuu $T \subset T_0$ lehtede kokku liitmisel. Sobivate lehtede kokku liitmisel kasutatakse *cost*

complexity pruning/weakest link pruning meetodit. Meetodi käigus leitakse parameetri $\alpha, \alpha \geq 0$, väärtuste korral alampuud T_α nii, et minimeeritakse

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) + \alpha \cdot |T|,$$

kus $|T|$ tähistab lehtede arvu puus. Mida väiksem on α, seda sügavam on puu. Lehti saab liita seni kuni jõutakse juurtipuni välja.



Joonis 2: Otsustuspuu üleshituse näide

Joonisel 2 on toodud näide töös kasutatavast andmestikust ja sellele ehitatud otsustuspuust. Otsustuspuu treenimiseks kasutati rakendustarkvara *R* paketti *rpart* (Therneau, Atkinson ja Ripley, 2022) ning joonise tekitamiseks kasutati paketti *rpart.plot* (Milborrow, 2021). Kujutatud puul on 6 lehttippu ning 4 vahetippu, mille defineerimiseks kasutati 3 tunnust. Joonisel kujutatud kastid annavad informatsiooni, mis on tipu poolt prognoositud klass, mis on hospitaliseerimise tõenäosus ja vaatluste osakaal, mis sattusid tippu. Iga tipu juures on toodud tingimus, mille

alusel on andmed kaheks piirkonnaks jaotatud.

1.5 Otsustusmets

Üksik puu ei anna enamasti head tulemust. Kasutades mudelis mitut erinevat puud saame kokkuvõttes hea tulemuse. Üks selline ansambelmeetod (*ensemble method*) on otsustusmets. (James *et al.*, 2021, lk 340) Järgnev alapeatükk põhineb raamatul James *et al.* (2021, lk 340–345), kui ei ole märgitud teisiti.

Otsustusmets koosneb mitmest otsustuspuust. Erinevate puude defineerimiseks võetakse mudeldamiseks kasutatav andmestik ning moodustatakse B andmestikku *bootstrap*-meetodil taasvalikuga. Iga b-nda puu ($b = 1, 2, ..., B$) tipu defineerimisel võetakse b-ndast andmestikust $m \approx \sqrt{p}$, kus p tähistab kirjeldavate tunnuste arvu, juhuslikult valitud tunnust. Selekteeritud m kirjeldava tunnuse seast valitakse tunnus, mis määrab puu tipu. Otsustusmetsaga saab leida vaatlusele nii konkreetsesse klassi kuulumise prognoos kui ka tõenäosuse kuuluda mingisse klassi. Edaspidi nimetatakse kirjeldatud metsi klassifitseerimis- ja tõenäosusmetsaks. Klassifitseerimismetsa puhul leitakse uuele vaatlusele prognoos iga puu prognoositud klassi enamuse põhjal. Tõenäosusmetsa korral annab puu klassi k korral prognoosiks piirkonnas leiduvate klassi k vaatluste osakaalu ning metsa prognoos leitakse puude prognooside keskmisena.

Otsustusmetsa defineerimisel saab määrata mitmeid hüperparameetreid (vt tabel 1), millel on ka vastavalt kasutatavale tarkvarale omad vaikeväärtused. Antud töö käigus kasutatakse otsustusmetsa defineerimiseks rakendustarkvara R paketti *ranger* (Wright, Wager ja Probst, 2021), mille nimetused koos vaikeväärtustega on toodud tabelis 1. Fikseeritud vaikeväärtustega on võimalik saada häid tulemusi, kuid mudeli võimekust on võimalik hüperparameetrite seadistamisega parandada.

Probst, Wright ja Boulesteix (2019) toovad oma artiklis välja otsustusmetsa hüperparameetrid ning nende mõju kui vaikeväärtuseid muuta. Kõige olulisem sea-

Tabel 1: Otsustusmetsa hüperparameetrid funktsiooni *ranger* korral

Hüperparameeter	**Parameetri nimetus** (*ranger*)	**Vaikeväärtus**
Valikumäär	sample.fraction	1 ehk valim suurusega n
Tunnuste arv	mtry	$\sqrt{p}$
Lehe puhtuse hinnang	splitrule	Gini impurity
Minimaalne lehe suurus	min.node.size	10 (tõenäosusmets), 1 (klassifitseerimismets)
Puude arv metsas	num.trees	500
Valimi moodustamine tagasipanekuga või tagasipanekuta	replace	TRUE (tagasipanekuga)

distatav parameeter on puude arv ning selgub, et otsustusmetsa korral on mõistlik kasutada rohkelt puid. Samal ajal tuleb arvestada, et puude lisamisel metsa kulub rohkem aega mudeli arvutusele ning, et mingist hetkest puude lisamine enam mudeli võimekust ei tõsta. Artiklis tuuakse välja, et lisaks puude arvule on oluline seadistada ka valitavate tunnuste arvu.

Puu defineerimiseks valitakse m tunnust p kirjeldava tunnuse seast. Kui ülesandeks on leida klassifitseerimismets, siis kasutatakse tihti $m \approx \sqrt{p}$. Mida väiksem on valitavate tunnuste arv, seda rohkem defineerivad puu tippe tunnused, mis ei ole tugevalt seotud uuritava tunnusega. See võib tuua positiivset efekti kui sellised tunnused suudavad kirjeldada mõne alamgrupi, mis tugevalt seotud tunnuste korral võivad varju jääda. Mida suurem on valitavate tunnuste arv, seda kindlam on see, et valikusse satub ka vähemalt üks tugevalt seotud tunnus. Selline lähenemine on hea, kui andmestik koosneb enamasti tunnustest, mis on vähemolulised. (Probst, Wright ja Boulesteix, 2019)

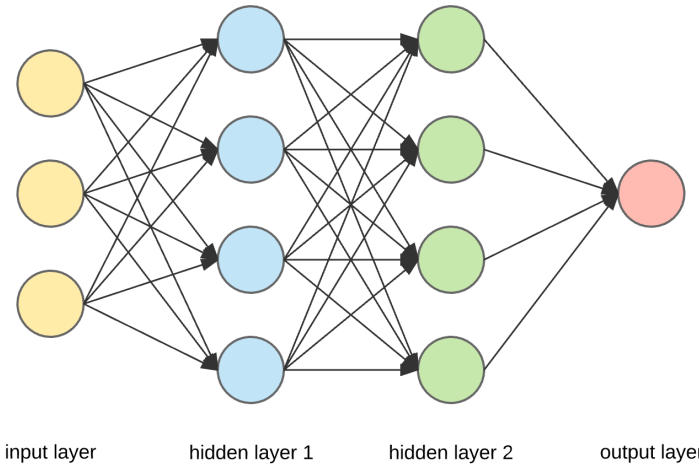
Lisaks toovad Probst, Wright ja Boulesteix (2019) välja, et valikumäära ja minimaalse lehe suuruse vaikeväärtuste muutmisel on väike mõju mudelile, kuid tasub proovimist. Valikumäär kirjeldab puu defineerimiseks kasutatava valimi suurust

ning minimaalse lehe suurus tähistab minimaalset vaatluste arvu lehes. Väikse valimi korral võtab mudeli arvutamine vähem aega. Samal ajal kaotatakse üksiku puu täpsuses, kuna puu defineerimisel kasutatakse vähem andmeid. Minimaalne lehe suurus määrab kaudselt ära, kui sügavaid puid defineeritakse. Väiksema lubatud arvu korral kirjeldatakse rohkem vahetippe. Minimaalse lehe suuruse kasvades väheneb mudeli arvutusele kuluv aeg.

1.6 Närvivõrgud

Järgnevas alapeatükis kirjeldatakse sügava pärilevivõrgu (*deep feedforward network*) metoodikat, mida lisaks nimetatakse veel mitmekihiliseks närvivõrguks (*multilayer perceptron*) ja pärilevi närvivõrguks (*feedforward neural networks*). Alapeatükk põhineb raamatutel Goodfellow, Bengio ja Courville (2016, ptk 4, 6–8), Bishop (2006, lk 227–246, 256–261) ja James *et al.* (2021, lk 403–411, 434–439), kui ei ole märgitud teisiti.

Pärilevivõrk koosneb sisendkihist (*input layer*), peidetud kihtidest (*hidden layer*) ja väljundkihist (*output layer*). Joonisel 3 on kujutatud pärilevi närvivõrk, mis



Joonis 3: Kahe peidetud kihiga pärilevi närvivõrk (Dertat, 2017)

koosneb sisendkihist, kahest peidetud kihist ning väljundkihist. Kihid koosnevad

sõlmedest ning kihtide sõlmed (*nodes*) on omavahel seotud joonte abil, mis tähistavad kaale (*weight parameters*). Pärilevivõrgu puhul liiguvad andmed mööda võrku edasi ning mudeli prognooside kohta tagasisidet ei anta. Kasutatud peidetud kihtide arv iseloomustab mudeli sügavust (*depth*) ning sõlmede arv peidetud kihis iseloomustab mudeli laiust (*width*).

Binaarse tunnuse $Y_i \in \{0, 1\}$ korral on soov, et närvivõrk prognoosiks $P(Y_i = 1)$. Selleks, et väljund kuuluks vahemikku $(0, 1)$ kasutatakse väljundkihis aktivatsioonifunktsioonina sigmoidfunktsiooni:

$$g(z) = \sigma(z) = \frac{1}{1 + \exp(-z)}.$$

Järgmisena vaadeldakse kahekihilist pärilevi närvivõrku, mis koosneb ühest peidetud kihist ning kasutab väljundkihis sigmoid-funktsiooni. Sisendkihi suurus on p ning sisendiks antakse vaatlus X_i. Esmalt leitakse lineaarkombinatsioonid:

$$a_l = \sum_{j=1}^{p} w_{lj}^{(1)} \cdot X_{ij} + w_{l0}^{(1)},$$

kus $l = 1, 2, ..., M$ ning M tähistab sõlmede arvu peidetud kihis. Parameetrid $w_{lj}^{(1)}$ tähistavad kaale (*weights*) ja $w_{l0}^{(1)}$ tähistab vabaliiget (*bias*) esimeses kihis. Seejärel teisendatakse väärtused a_l aktivatsioonifunktsiooniga:

$$z_l = h(a_l).$$

Saadud väärtused z_l on väljundid peidetud kihist. Mitme peidetud kihi korral moodustatakse saadud peidetud kihi väärtustega uued lineaarkombinatsioonid. Hetkel moodustatakse üks lineaarkombinatsioon peidetud kihi väljunditest z_l:

$$a = \sum_{l=1}^{M} w_l^{(2)} \cdot z_l + w_0^{(2)},$$

millele rakendatakse sigmoid-funktsiooni ning saadakse prognoos:

$$\hat{Y}_i = \hat{y}(X_i, \mathbf{w}) = \sigma(\sum_{l=1}^{M} w_l^{(2)} \cdot h(\sum_{j=1}^{p} w_{lj}^{(1)} X_{ij} + w_{l0}^{(1)}) + w_0^{(2)}).$$

Andmete liikumist sisendkihist väliskihi suunas, nagu eelnevalt kirjeldati, nimetatakse pärileviks (*forward propagation*).

Närvivõrgu ülesehitust iseloomustab peidetud kihtide arv ja kihi suurused ning kasutatud aktivatsioonifunktsioonid. Peidetud kihis kasutatakse aktivatsioonifunktsiooni vaikeväärtusena mittenegatiivset lineaarfunktsiooni ReLu (*Rectified Linear Unit*)

$$g(z) = \max\{0, z\} = \begin{cases} 0, \text{kui } z < 0, \\ z, \text{kui } z \geq 0. \end{cases}$$

Binaarse klassifitseerimisülesande korral kasutatakse närvivõrkude treenimisel kaofunktsioonina ristentroopiat (*cross-entropy*)

$$E(\mathbf{w}) = \sum_{i=1}^{n} E_i(\mathbf{w}) = -\sum_{i=1}^{n} \{Y_i \cdot \ln(\hat{Y}_i) + (1 - Y_i) \cdot \ln(1 - \hat{Y}_i)\},$$

kus Y_i tähistab uuritava tunnuse väärtust i-nda vaatluse korral ning $\hat{Y}_i$ tähistab mudeli prognoosi i-ndale vaatlusele. Treenimise protsessis soovitakse leida kaalud $\mathbf{w}$ nii, et kaofunktsioon oleks minimaalne. Sobivad kaalud leitakse iteratiivselt alustades juhuslikult valitud kaaludega $\mathbf{w}^{(0)}$. Igal sammul t leitakse sobiv muudatuste vektor. Kaalude muudatuste sobiv suund leitakse gradiendi $\nabla E(\mathbf{w})$ kaudu, mis viitab kaofunktsiooni suunale, kus tõus on suurim. Seega leitakse kaalud

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla E(\mathbf{w}^{(t)}),$$

kus $\eta > 0$ on õpisamm (*learning rate*). Kirjeldatud meetodit nimetatakse gradientlaskumiseks (*gradient descent*).

Meetodi kasutamiseks on vaja leida $\nabla E(\mathbf{w})$, milleks kasutatakse tagasilevi (*back-propagation*). Vaja on leida tuletised kaalude $\mathbf{w}$ järgi. Hetkel käsitletava näite korral leitakse osatuletised esimese kihi kaalude suhtes $\frac{\partial E_i}{\partial w_{lj}^{(1)}}$ ja teise kihi kaalude suhtes $\frac{\partial E_i}{\partial w_l^{(2)}}$:

$$\frac{\partial E_i}{\partial w_l^{(2)}} = \frac{\partial E_i}{\partial \hat{Y}_i} \cdot \frac{\partial \hat{Y}_i}{\partial a} \cdot \frac{\partial a}{\partial w_l^{(2)}},$$

$$\frac{\partial E_i}{\partial w_{lj}^{(1)}} = \frac{\partial E_i}{\partial \hat{Y}_i} \cdot \frac{\partial \hat{Y}_i}{\partial a} \cdot \frac{\partial a}{\partial z_l} \cdot \frac{\partial z_l}{\partial a_l} \cdot \frac{\partial a_l}{\partial w_{lj}^{(1)}}.$$

Kui gradiendi leidmiseks kasutatakse tervet andmestikku, siis nimetatakse meetodit *batch gradient method*. Enamasti ei kasutata praktikas arvutamiseks kogu treeningandmestikku, vaid ainult mingit osa. Sellisel juhul on tegemist miniplokk-stohhastilise meetodiga (*minibatch stochastic method*). Miniploki suurusena kasutatakse sageli väärtuseid 32, 64, 128, 256, mis annavad mudeli jooksutamisel kiiremad tulemused. Lisaks määratakse enne treenimist, mitu korda treeningprotsessis andmestik läbitakse. Andmestiku läbimist nimetatakse epohhiks (*epoch*). Epohhi jooksul kasutatakse $\frac{n}{\text{miniploki suurus}}$ miniplokki. Stohhastilise gradientlaskumise korral leitakse ühe epohhi korral miniplokkides arvutatud gradientide keskmine. Rakendustarkvara *Python* pakett Keras (Abadi *et al.*, 2015) kasutab treeningprotsessis stohhastilise gradiendilaskumise algoritmi Adam (Kingma ja Ba, 2017).

Sobivate kaalude leidmiseks saab kaofunktsioonile lisada kaaludele karistusliikme (*penalty term*), mille eesmärk on vähendada viga, mida mudel teeb uute andmete korral ehk vältida ülesobitamist. Kirjeldatud protsessi nimetatakse regulariseerimiseks (*regularization*). Kaalude kahandamiseks (*weight decay*) nulli lähedale saab kasutada kantregularisatsiooni:

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w},$$

kus $\lambda \in [0, \infty)$ tähistab regulariseerimise parameetrit. Kui $\lambda = 0$, siis regulariseе-

rimist ei toimu.

Närvivõrkude puhul on soovitatav seadistada peidetud kihtide arvu, peidetud kihtide suuruseid, regulariseerimisparameetri väärtust, miniploki suurust ja epohhide arvu. Epohhide arvu määramine on oluline, sest andmestikku mitmeid kordi läbides hakkab mudel ühel hetkel ülehindama. Epohhide arvu määramiseks saab kasutada ka funktsiooni *EarlyStopping*, mis treenimisel jälgib, mis hetkest kaofunktsioon enam ei vähene ning lõpetab seejärel protsessi (Keras, 2022).

1.7 Mudeli võimekuse hindamine

Järgnev alapeatükk põhineb raamatul Murphy (2012, lk 183–185), kui ei ole märgitud teisiti.

Masinõppe meetodite eesmärk on õppida olemasolevatelt andmetelt ning teha võimalikult vähe vigu uute andmete korral. Selleks, et saada aimu, kuidas saab mudel uute andmete korral hakkama, jagatakse algne andmestik treening- ja testandmeteks (vt alapeatükk 1.1). Mudeli abil saab prognoosida tulemused testandmetel ning seejärel uurida, kui täpselt saab mudel hakkama. Saadud tulemused kantakse maatriksisse (vt tabel 2), kust saab välja lugeda erinevaid mudeli võimekuse (*model performance*) näitajaid. Kirjeldatud maatriksit tuntakse kui *confusion matrix*.

Kirjeldatud tabelist saame arvutada erinevaid mõõdikuid:

- Õigsus (*accuracy*) $= \frac{TN+TP}{TN+FP+FN+TP}$
- Tundlikkus (*sensitivity, true positive rate, recall*)

$$P(\hat{Y} = 1|Y = 1) = \frac{TP}{TP + FN}$$

Tabel 2: Klassifitseerija tulemused testandmetel

		Prognoos		
		0	1	
Tegelik	0	TN	FP	TN+FP
	1	FN	TP	FN+TP

* TN tähistab õige negatiivsete tulemuste arvu (*true negative*),
TP tähistab õige positiivsete tulemuste arvu (*true positive*),
FP tähistab valepositiivsete tulemuste arvu (*false positive*),
FN tähistab valenegatiivsete tulemuste arvu (*false negative*).

- Spetsiifilisus (*specifity, true negative rate*)

$$P(\hat{Y} = 0|Y = 0) = \frac{TN}{FP + TN}$$

- Valepositiivsuse määr (*false positive rate*)

$$P(\hat{Y} = 1|Y = 0) = \frac{FP}{FP + TN}$$

- Valenegatiivsuse määr (*false negative rate*)

$$P(\hat{Y} = 0|Y = 1) = \frac{FN}{FN + TP}$$

- Täpsus (*precision, positive predictive value*) – näitab, kui suur hulk positiivsetest prognoosidest on tegelikult positiivsed

$$P(Y = 1|\hat{Y} = 1) = \frac{TP}{TP + FP}.$$

Klassifitseerimisülesannete korral on kõige populaarsem mõõdik õigsus. Mudeli võimekuse hindamiseks tuleb valida lähtuvalt eesmärgist sobiv mõõdik.

1.8 Tasakaalustamata andmed

Järgnev alapeatükk põhineb allikatel Ling ja Sheng (2010) ja Provost (2000), kui ei ole märgitud teisiti.

Tasakaalustamata andmed tähistavad olukorda, kus uuritava tunnuse Y võimalike väärtuste sagedused on erinevad. Binaarse tunnuse korral koosneb andmestik rohkelt esindatud klassist ehk enamusklassist (*majority class*) ning vähe esindatud klassist ehk vähemusklassist(*minority class*). Tasakaalustamata andmete korral võib andmestik koosneda näiteks 99% enamusklassi ja 1% vähemusklassi vaatlustest. Enamasti pakub huvi just vähemusklassi prognoosimine. Enamus klassifitseerijaid treenitakse nii, et õigsus oleks võimalikult madal, mis väga tasakaalustamata andmete korral saavutatakse ka ilma vähemusklassile tähelepanu pööramata.

Tasakaalustamata andmete probleem seisneb selles, et vead, mida mudel prognoosimisel teeb, ei ole tegelikult võrdväärsed. Lihtne näide on mõne raske haiguse võimalike põdejate tuvastamine, et suunata nad täpsemale uurimisele ja jälgimisele. Enamasti on raske haigusega patsiente andmetes vähe, kuid soov on uutel inimestel tuvastada raske haigus varakult. Nüüd patsiente klassifitseerides terveteks ja haigeteks on selge, et haige inimese klassifitseerimine terveks on palju kallim viga kui terve inimese klassifitseerimine haigeks.

Selgub, et tehes mudelis muudatusi enne või peale treenimist, saab juba tuntuid klassifitseerimise meetodeid kasutada tasakaalustamata andmete mudeldamiseks. Antud töö käigus tutvutakse kahe meetodiga:

1. lävendi meetod (*thresholding*),
2. valikumeetod (*sampling*).

1.8.1 Lävendimeetod

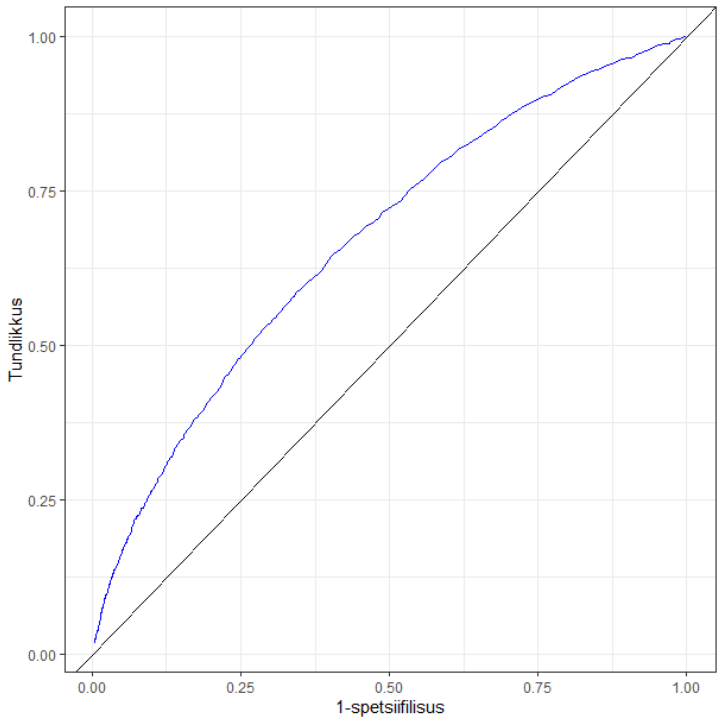
Järgnev alapeatükk põhineb raamatul Murphy (2012, lk 183–184), kui ei ole märgitud teisiti.

Defineeritud klassifitseerija ei pea välja andma klassi kuuluvuse hinnanguid, vaid võib hinnata ka, kui kindlalt vaatlus kuulub huvipakkuvasse klassi. Binaarse tunnuse korral on vaatluse all enamasti $Y_i = 1$. Mudeli poolt saadav hinnang $f(X_i)$ ei pea olema tõenäosus $p(Y_i = 1|X_i)$, vaid võib olla ka hinnang, mis on monotoonselt seotud tõenäosusega, et $Y_i = 1$. Saadud hinnangu $f(X_i)$ abil saab määrata vaatlused klassidesse. Fikseerides mingi lävendi τ, siis

$$\hat{Y}_i = \mathrm{I}(f(X_i) > \tau).$$

Nüüd saab tulemused kanda tabelisse 2. Kasutatud meetodit nimetatakse lävendi meetodiks (*threshold method*). Provost (2000) toob oma artiklis välja, et tasakaalustamata andmete korral võib piisata tavalise klassifitseerimismeetodi korral lävendi muutmisest.

Sobiva lävendi τ valikuks saab kasutada ROC-kõverat (*receiver operating characteristic*) (vt joonis 4). Kõigepealt leitakse *tundlikkus* ja 1− *spetsiifilisus* väärtused erinevate lävendite korral. Kui lävend $\tau = 1$, siis klassifitseeritakse kõik testandmestiku vaatlused klassi 0. Järelikult mõlemad näitajad, *tundlikkus* ja 1−*spetsiifilisus*, on võrdsed nulliga. Kui lävend $\tau = 0$, siis klassifitseeritakse kõik vaatlused klassi 1. Järelikult mõlemad näitajad on võrdsed ühega. Kui kõver langeb kokku diagonaaljoonega, siis kirjeldatud näitajad on võrdsed. Selline seisund tähendab, et mudel ei suuda klasse eristada ning käitub juhuslikult. Soov on leida lävend τ, mis annab määratud kindluse prognooside osas. Näiteks tahetakse võimalikult kõrget *tundlikkuse* määra. Enamasti on soov leida lävend, mille korral *tundlikkus* ja *spetsiifilisus* oleks samaaegselt võimalikult kõrged ehk mudel suudaks kaks klassi omavahel täiuslikult eristada. Sellist kõverat iseloomustaks olukord, kui joonisel 4



Joonis 4: ROC-kõvera näide logistilise regressioon mudeli korral

kujutatud kõver jõuaks üles vasakule nurka. Sellisel juhul oleks nii *tundlikkus* kui ka *spetsiifilisus* võrdsed ühega.

Joonisel 4 kujutatud kõver ei suuda perfektselt klasse eristada, kuid on parem kui juhuslik klassifitseerimine. Lisaks ROC-kõverale kasutatakse lävendi leidmiseks ka teisi mõõdikuid, nt *precision-recall* kõverat. Lävendi valimine oleneb ülesandest ja sellest, kumb viga on kallim.

1.8.2 Valikumeetod

Järgnev alapeatükk põhineb allikal Provost (2000), kui ei ole märgitud teisiti.

Teine meetod on andmete töötlemine enne mudeldamist. Võimalus on klasside jaotused võrdsustada ülevaliku (*upsampling*) ja alavaliku (*downsampling*) abil. Lisaks on võimalik lisada treeningandmetele kaalud, mida arvestatakse mudeli treenimisel.

Valimi muutmise eesmärk on tehislikult viia uuritavad klassid tasakaalu. **Alavaliku** korral jäetakse mingi hulk enamusklassi vaatluseid kõrvale. **Ülevaliku** korral paljundatakse olemasolevaid vähemusklassi vaatluseid. Nende meetodite miinus on see, et alavaliku korral kaotatakse mingi osa informatsiooni ja ülevaliku puhul ei lisata andmestikku uut informatsiooni.

Vaatlustele lisatud kaale kasutavad erinevad meetodid erinevalt. Otsustusmetsa korral, kasutades funtsiooni *ranger*, on võimalik määrata kaks parameetrit: *class.weights* ja *case.weights*. Esimese korral antakse klassidele kaalud, mida kasutatakse vahetippude defineerimisel. Lisaks kasutatakse klassifitseerimismetsa treenimisel kaale lehetipus prognoosi leidmisel ehk arvutatakse kaalutud enamus (*weighted majority vote*). Klassidele saab lisada kaale klassifitseerimis- või tõenäosusmetsa treenimisel. Teine parameeter on vaatluste kaalud, mida kasutatakse *bootstrap*-meetodiga valimite koostamisel. Suurema kaaluga vaatlused võetakse suurema tõenäosusega valimisse. (Wright, Wager ja Probst, 2021)

2 Andmed

Järgnevas peatükis antakse ülevaade riskipatsientide valimi moodustamisest ning analüüsi jaoks kasutatud andmetest. Andmete töötlus teostatakse rakendustarkvaraga *R* ning kasutatakse paketti *tidyverse* (Wickham *et al.*, 2019).

2.1 Riskipatsiendid

Eesti Haigekassa ja Maailmapank viisid 2015. aastal läbi uuringu "Ravi terviklik käsitlus ja osapoolte koostöö Eesti tervishoiusüsteemis", mille käigus keskenduti krooniliste haigustega patsientidele ning nende haiguste ennetusele ja raviteekonnale (Eesti Haigekassa ja Maailmapanga Grupp, 2015). Uuringu jätkuna viidi aastatel 2016–2017 Eestis läbi ravi juhtimise pilootprojekt (Maailmapanga Grupp, 2017). Uuringuga leiti, et statsionaarses aktiivravis on palju välditavaid juhtumeid ning toodi välja, et neid hospitaliseerimisi on võimalik vältida perearsti ennetavate tegevustega. Lisaks väideti, et üks oluline välditavate hospitaliseerimiste allikas on mitmete krooniliste haigustega patsiendid, kes ei käi piisavalt sageli arstlikul jälgimisel, mistõttu nende krooniline haigus võib ägeneda ja viia haiglaravi vajaduseni. Selliseid patsiente nimetatakse riskipatsientideks. Antud töö eesmärk on luua riskipatsientidele hospitaliseerimise riski mudel. Soov on, et mudel prognoosiks eelmise aasta raviarvete andmete põhjal, kas patsiendil on risk järgmisel aastal välditavaks hospitaliseerimiseks. Leitava mudeliga soovitakse riskipatsiendid jagada kriitilisteks ja vähem kriitilisteks juhtumiteks. Antud töö raames loetakse välditavaks hospitaliseerimiseks sellist haiglaravi juhtu, mis ei ole põhjustatud nakkushaigusest, kasvajast, rasedusest, sünnitusest ega sünnitusjärgsest perioodist, vigasdusest ega muust õnnetusest (RHK-10 koodide vahemikud A00–D89, O00–P96, S00–T98, samuti ei tohi hospitaliseerimisel olla välispõhjust, aga lubatud on ravimitest põhjustatud välispõhjused (RHK-10 koodide vahemikud Y40–Y84)). Käesoleva töö käigus mõeldakse hospitaliseerimise all eelnevalt kirjeldatud välditavaid hospitali-

seerimise juhtumeid.

Antud analüüsiks kasutatakse valdavalt Eesti Haigekassa andmebaasis leiduvaid andmeid. Kasutades Maailmapanga raportis (Maailmapanga Grupp, 2017) toodud riskipatsientide väljasõelumise algoritmi, moodustatakse riskipatsientide valim järgnevalt. Esmalt võetakse vaatluse alla kõik isikud, kellel esines perioodil 01.01.2017–31.12.2018 põhi- või kaasuva diagnoosina[1]

- hüpertensioon (I10–I15),
- hüperlipideemia (E78) või
- diabeet (E11–E14).

Seejärel vaadeldakse perioodil 01.01.2015–01.01.2016 isikuid, kes ei sattunud eelmisse nimekirja, kuid kellel esines põhi- või kaasuva diagnoosina hüpertensioon, hüperlipideemia või diabeet. Nendest võetakse vaatluse alla patsiendid, kellele perioodil 01.01.2017–31.12.2018 ei esitatud rohkem kui 4 perearstiarvet. See tähendab, et kõrvale jäetakse patsiendid, kes olid juba perearsti jälgimise all.

Hetkel valimisse kuuluvatest isikutest jäetakse kõrvale need, kellel perioodil 01.07.–31.12.2018 diagnoositi vähk (need patsiendid on tõenäoliselt juba arstliku jälgimise all). Lisaks jäetakse kõrvale patsiendid, kellel samal perioodil diagnoositi skisofreenia, neerupuudulikkus ja kes käivad dialüüsil, kaasasündinud väärarendid või harvaesinevad haigused. Veel jäetakse kõrvale isikud, kellel sel perioodil diagnoositi kaasuvana 7 või enam haigust järgnevatest: aneemia, kilpnäärme haigusseisund, rasvumus, krooniline südamepuudulikkus jt (vt lisa 1). Seejärel jäetakse vaatluse alla ainult need patsiendid, kellel esineb vähemalt üks diagnoos järgnevatest:

1. peaaju transitoorse isheemia atakid ja selle sarnased sündroomid (G45),

[1]Diagnoosid ja nende koodid vastavad Rahvusvahelise Haiguste Klassifikatsiooni 10. versioonile (RHK-10).

2. südame isheemiatõved (I20–I25),

3. kodade virvendus ja laperdus (I48),

4. kõrgvererõhkhaigused (I11.0, I13.0, I13.2) või muud südamehaigused (I50.0, I50.1, I50.9),

5. alumiste hingamisteede kroonilised haigused (J40–J47),

kuid ei esine rohkem kui kaks diagnoosi 1.–4. seast. Viimasena jäetakse kõrvale isikud, kellel esineb rohkem kui 1 vaimse häire diagnoosi (vt lisa 2). Lõplikkusse valimisse kaasatakse isikud, kes olid elus 31.12.2018 seisuga ning kes 2018. aasta alguse seisuga olid vanemad kui 25 eluaastat. Valimi moodustamise kriteeriumid on valitud põhimõttel, et riskipatsientide hulka jääks kroonilise haigusega patsiendid, kelle puhul ravisekkumine võib tuua tervisele kasu ning kellest teavitamine on perearstile seetõttu vajalik.

2.2 Tunnused

Mudeli välja töötamiseks võetakse valimisse sattunud patsientide kohta andmed raviarvetelt ja retseptidelt, mis esitati 2018. aasta jooksul. Kasutatud andmed kirjeldavad patsiendi terviseseisundit hetkeks 31.12.2018. Andmed hospitaliseerimiste kohta võetakse aastast 2019. Uuritav tunnus kirjeldab, kas aastal 2019 toimus hospitaliseerimist või mitte. Terviseseisundit kirjeldatakse erinevate tunnustega.

1. Üldised andmed: *Vanus*, *Sugu*.

2. Elukohta iseloomustavad tunnused:

    - *kaugus_m* – elukoha kaugus meetrites lähimast haiglast,
    - *ElabYksi* – binaarne tunnus, kas patsient elab üksi,
    - *vaesus* – suhtelise vaesuse määr elukoha maakonnas aastal 2018.

3. Haiglas veedetud voodipäevade arv:

    - *VoodipäeviOendusabi* – raviarved teenusetüübiga 18 (Eesti Haigekassa, 2021, alapeatükk 1.5.5),
    - *VoodipäeviTaastusravi* – raviarved teenusetüübiga 15,
    - *VoodipäeviIntensiivravi* – raviarved teenustega (Riigikantselei ja Justiitsministeerium, 2022a) 2070, 2071, 2072 või 2073,
    - *VoodipäeviMuu* – voodipäevade arv, mis ei ole seotud õendusabi, taastusravi ega intensiivraviga,
    - *Voodipäevi* – voodipäevade arv kokku.

4. Operatsioonide arv raviarvetel:

    - *Operatsioone* – tervishoiuteenuste arv liigiga *Operaatsioonid*,
    - *Operatsioon1kuni3h* – NCSP[2] koodiga ZXE10 (Tervise ja Heaolu Infosüsteemide Keskus, 2022) kirurgilise protseduuri kordade arv,
    - *OperatsioonHingamiselunditel* – kirurgiliste protseduuride kordade arv, mille NCSP kood algab tähega "G", ,
    - *OperatsioonErakorraline* – NCSP koodiga ZXD00 kirurgilise protseduuri kordade arv.

5. Esitatud raviarvete arv:

    - *ValtimatuAbiArveid* – raviarved, mis on märgitud vältimatuks,
    - *ArveidTeenuseTyyp2* – statsionaarne abi,
    - *ArveidTeenuseTyyp15* – statsionaarne taastusravi,
    - *ArveidTeenuseTyyp16* – ambulatoorne taastusravi,
    - *ArveidTeenuseTyyp18* – iseseisev stastionaarne õendusabi,

[2]NOMESCO kirurgiliste protseduuride klassifikatsioon.

- *ArveidTeenuseTyyp20* – koduõenduse,
- *AmbArveidPerearst* – perearsti poolt kirjutatud ambulatoorsed raviarved,
- *AmbArveidEriarst* – eriarsti poolt kirjutatud ambulatoorsed raviarved,
- *EMOvisiit* – erakorralise meditsiini raviarvete arv,
- *sots_ seisund* – raviarvete arv, millele on märgitud teenused 9142, 3026 või 3027 (vt lisa 3).

6. Raviteenuste arv arvetel kokku:
   - *Raviteenus66100*, *Raviteenus66101*, *Raviteenus66102* jt (vt lisa 3).
7. Määratud diagnoosid põhi- või kaasuva haigusena:
   - *Diagnoos krooniline südamepuudulikkus*, *Diagnoos astma*, *Diagnoos ärevushäire* jt (vt lisa 1) – binaarne tunnus, kas esineb diagnoos,
   - *DiagnoosiGrArv* – diagnooside arv kokku,
8. Väljakirjutatud retseptid:
   - *RetsepteKokku* – retseptide arv kokku,
   - *RetseptideHindKokku* – retseptide hind kokku,
   - *OmaosalusKokku* – omaosalus kokku,
   - *ValjaOstetudRetsepte* – retseptide arv staatusega "müüdud",
   - *KirjutatudRetsepteA01*, *KirjutatudRetsepteA02*, *KirjutatudRetsepteA03* jt (vt lisa 4) – välja kirjutatud retseptide arv toimeaine ATC (2. taseme) koodi kaupa (Ravimiamet, 2022).
9. Patsiendi kindlustusliik (Eesti Haigekassa, 2022a):

- *KindlustuseLiikTootu* – binaarne tunnus, kas omab üht kindlustust loetelust: töötu abiraha saav isik, töötuskindlustushüvitise saaja, töötu koolitus, töötu tööpraktika, tööharjutuses osaleja, töötu,
- *KindlustuseLiikVanaduspensionar* – binaarne tunnus, kas omab üht kindlustust loetelust: vanaduspensonär, Eesti pensionär teises EL liikmesriigis, välislepingu alusel kindlustatud pensionär, Vene Föderatsiooni sõjaväepensionär, EL pensionär, pensionär,
- *KindlustuseLiikSotsiaaltoetuseSaaja* – binaarne tunnus, kas omab üht kindlustust loetelust: sotsiaaltoetust saav isik, toitjakaotuspensionär,
- *KindlustuseLiikToovometu* – binaarne tunnus, kas omab üht kindlustust loetelust: töövõimetuspensionär, osalise või puuduva töövõimega isik,
- *KindlustuseLiikMuu* – binaarne tunnus, kas omab kindlustust, mis ei kuulu eelnevate kindlustuste hulka.

10. *PerearstiKvaliteediSkoorI_II* – perearsti kvaliteedisüsteemi (Eesti Haigekassa, 2022b) alusel patsiendi perearstile määratud punktid aastal 2017.

Tunnuse *kaugus_m* leidmiseks kasutatakse Maa-ameti Geoportaali (Maa-amet, 2021). Lähim haigla leitakse haiglate loetellu (Riigikantselei ja Justiitsministeerium, 2022b) kuuluvate haiglate seast[3]. Suhtelise vaesuse määr põhineb Statistikaameti andmetel (Statistikaamet, 2022). Eesti Haigekassa loeb vältimatuks abiks tervishoiuteenust, mille edasilükkamine või andmata jätmine võib põhjustada abivajaja surma või püsiva tervisekahjustuse (Eesti Haigekassa, 2022c).

2.3 Andmete töötlus

Huvipakkuvaid tunnuseid tuli kokku 213 ning riskipatsiente sattus valimisse 97 717. Esmalt otsitakse üles tunnused, mis on omavahel väga tugevalt korreleeritud. Sel-

[3]Välja arvatud SA Haapsalu Neuroloogiline Rehabilitatsioonikeskus.

gub, et tugevalt on seotud *ArveidTeenuseTyyp15* ja *VoodipaeviTaastusravi*, *ArveidTeenuseTyyp18* ja *VoodipaeviOendusabi* ning *RetsepteKokku* ja *ValjaOstetudRetsepte*. Nende paaride korral moodustatakse uued tunnused:

1. *Arve_ VPT* tähistab arvete arvu voodipäeva kohta taastusravil,
2. *Arve_ VP_ OA* tähistab arvete arvu voodipäeva kohta õendusabis.

   Tunnused leitakse voodipäevade arvu jagamisel arvete arvuga ning ümardamisel täisarvuks.
3. *Valja Ostmata* tähistab välja ostmata retseptide osakaalu.

   Tunnus leitakse: $1 - \frac{Valja\ Ostetud\ Retsepte}{Retsepte\ Kokku}$.

Tunnuseid uurides selgub, et tunnust *Raviteenus66138* ei esine ühelgi patsiendil ning seega jäetakse see kõrvale. Lisaks leidub tunnuseid, mida on vähe esindatud (vaatlusi alla 1 000). Sellised tunnused otsustatakse analüüsist välja jätta: *Operatsioon Hingamiselunditel*, *Kindlustuse Liik Sotsiaaltoetuse Saaja*, *Diagnoos ainete sõltuvus*, *Diagnoos alkoholi kuritarvitamine*, *Diagnoos kõne ja keele spetsiifilised arenguhäired*, *Diagnoos hüpotensioon*, ja *Diagnoos söömishäired*. Veel jäeti kõrvale kirjutatud retseptid toimeainetega V01, V03, V07, B02, B05, H01, H04, J02, J04, S02, M09, N01, R02, R05, R07 ja T01 ning raviteenused 66114, 66115, 66116, 66119, 66120, 66121, 66122, 66124, 66125, 66126, 66128, 66130, 66131, 66132, 66133, 66136, 66139 ja 9117. Vähe esindatud retseptide seast leitakse sarnase toimeainega grupid, mis liidetakse omavahel kokku, et tekitada uued tunnused:

1. *Kirjutatud Retsepte A* tähistab retseptide arvu toimeainetega A01, A04, A05, A07, A08, A09, A12, A14 ja A16.
2. *Kirjutatud Retsepte D* tähistab retseptide arvu toimeainetega D02, D03, D04, D05, D08, D10 ja D11.
3. *Kirjutatud Retsepte G* tähistab retseptide arvu toimeainetega G01 ja G02.

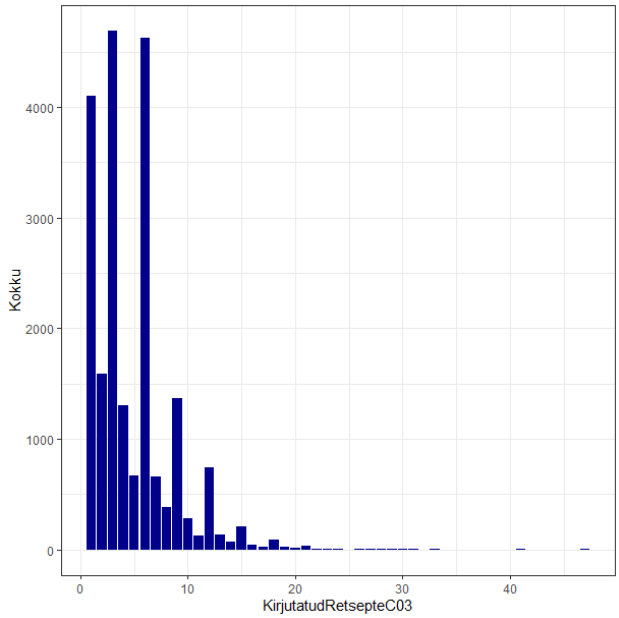
4. *Kirjutatud Retsepte L* tähistab retseptide arvu toimeainetega L01, L02, L03 ja L04.

5. *Kirjutatud Retsepte P* tähistab retseptide arvu toimeainetega P01, P02 ja P03.

Enamik tunnuseid andmetes on diskreetsed tunnused. Neid lähemalt uurides selgub, et enamik vaatlustest on väärtusega "0" ning teistest suuremate väärtustega tasemetel esineb üksikuid vaatluseid. Teistest erinevate tasemete korral otsustatakse tunnuse väärtused lõigata tasemeni, mida kirjeldab üle 1 000 vaatluse. Tunnuste teisendamise käigus muudetakse mõned diskreetsed tunnused binaarseks. Näiteks tunnust *Operatsioon 1 kuni 3h* ei esine 97%-il vaatlustest. Väärtus "1" esineb 3 106 inimesel ning kõigest 287 isikul on väärtus suurem ühest. Seega tunnus *Operatsioon 1 kuni 3h* muudetakse binaarseks tunnuseks ehk kas patsiendile tehti vastav kirurgiline protseduur vaatlusperioodi jooksul või mitte.

Tunnuste puhul, millel on rohkem kui kaks piisavalt kirjeldatud taset, koondatakse viimase kirjeldava tasemeni (vähemalt 1 000 vaatlust). Näiteks tunnuse *Operatsioone* puhul on väärtusega "1" 8 716 vaatlust, väärtusega "2" 2 217 vaatlust ning rohkem kui kahe operatsiooniga 361 vaatlust. Teisendamise käigus koondatakse rohkem kui kahe operatsiooniga vaatlused tasemele "2". Teisendatakse kõik tunnused, millel esineb vähe kirjeldatud suuri tasemeid (vt lisa 5).

Mõned tunnused muudetakse ka kategooriliseks. Joonisel 5 võib näha, et tunnuse *Kirjutatud Retsepte C03* korral kirjutatakse retsepti välja rohkem väärtustega "3", "6", "9", "12". Tunnuse jaotuse põhjal otsustatakse muuta tunnus kategooriliseks võttes üheks kategooriaks vahemiku 1 kuni 3, 4 kuni 6 jne. Sarnast käitumist nähakse ka C01, C07, C08, C09, C10 toimeainetega retseptide puhul. Andmeid uurides jääb veel silma voodipäevade arvu kirjeldavad tunnused, millel on vähestel nullist erinev väärtus ning nullist suuremad väärtused ei ole täisarvud. Seetõttu otsustatakse tunnused muuta kategooriliseks. Lisaks muudetakse kategooriliseks ka uued



Joonis 5: Tunnuse *Kirjutatud Retsepte C03* nullist suuremate väärtuste jaotus

tunnused *Arve_ VPT* ja *Arve_ VP_ OA*. (vt lisa 6)

Pidevatest tunnustest korrigeeritakse tunnuste *Retseptide Hind Kokku* ja *Omaosalus Kokku* jaotuseid. Tunnuse *Retseptide Hind Kokku* korral jääb silma, et pooled vaatlustest on väärtusega kuni 263,84. Samal ajal maksimaalne võimalik väärtus on rohkem kui 70 000. Samuti nähakse sarnast käitumist ka tunnuse *Omaosalus Kokku* puhul, kus leidub üksikuid väga suuri vaatlusi. Mõlema tunnuse puhul koondatakse väga suured väärtused väiksema tasemeni. Tunnuse *Retseptide Hind Kokku* puhul lõigatakse tunnused tasemeni 3 000 ja tunnuse *Omaosalus Kokku* korral tasemeni 700.

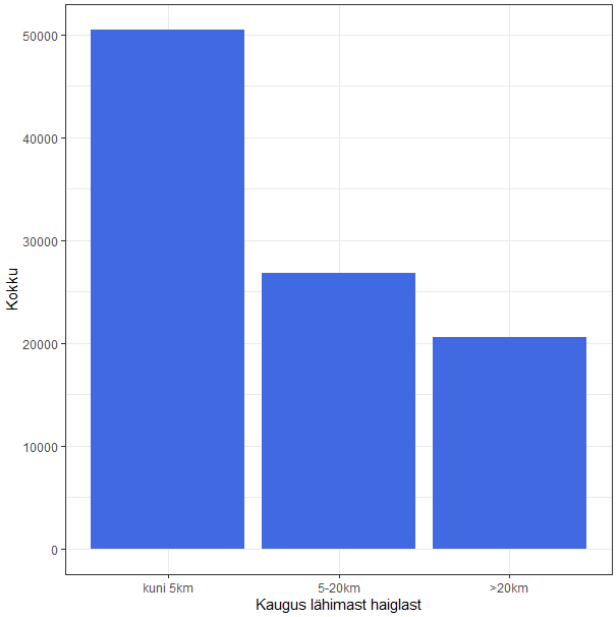
2.4 Andmete kirjeldus

Peale tunnuste töötlust jääb alles 146 kirjeldavat tunnust. Valimisse kuuluvatest 97 717 riskipatsiendist 39% mehed ning 61% naised. Riskipatsiendi keskmine vanus on 69 eluaastat. Samal ajal valimisse sattunud naiste keskmine vanus on 71,7

eluaastat ja meestel 65,4 eluaastat. Valimisse kuuluvatest isikutest 75%-il on kindlustuse liik *Vanaduspensionär*.

Tabel 3: Andmete kokkuvõte

Sugu	**Hospitaliseerimine aastal 2019**	**Juhtumeid**	**Osakaal**	**Keskmine vanus**
Naine	Ei	50 768	(52%)	71,3
Naine	Jah	8 627	(9%)	73,9
Mees	Ei	31 830	(33%)	64,8
Mees	Jah	6 492	(6%)	67,9

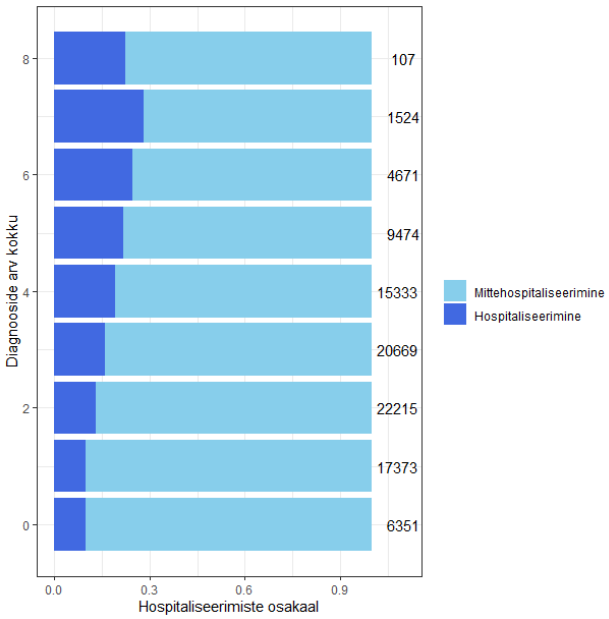


Joonis 6: Kaugus lähimast haiglast

Tabelist 3 võib näha, et nii naiste kui ka meeste puhul on hospitaliseeritute seas keskmine vanus kõrgem kui mittehospitaliseeritutel. Samuti jääb silma, et riskipatsientidest meeste keskmine vanus on madalam kui naistel. Hospitaliseerimisi esineb 15%-il riskipatsientidest ning meeste hulgas on hospitaliseeritute osakaal suurem.

Riskipatsientidest elab lähim 54 meetri kaugusel lähimast haiglast ning kõige kaugem asub 70 kilomeetri kaugusel. Jooniselt 6 on näha, et ligikaudu pooled riski-

patsientidest elavad kuni 5 kilomeetri kaugusel. Riskipatsiendi elukoha keskmine kaugus lähimast haiglast on 10,3 kilomeetrit.



Joonis 7: Hospitaliseeritute osakaal erineva diagnooside arvuga patsientide seas

Joonisel 7 on toodud määratud huvipakkuvate diagnooside (vt lisa 1) arv kokku. Enamikule patsientidele on määratud 1 kuni 4 diagnoosi. Patsientidele, kellele on määratud 8 diagnoosi, on perioodil 01.01.2017–31.06.2018 diagnoositud vähk. Graafikult on näha, et diagnooside arvu kasvades suureneb ka hospitaliseeritute osakaal. Seitsme diagnoosi korral on hospitaliseeritute osakaal 28%. Kõige rohkem on riskipatsientidele vaatlusperioodil määratud diagnoosi krooniline südamepuudulikkus (46%), südame isheemiatõved (30%) ja kodade virvendus ja laperdus (23%). Kõige vähem esineb somatoformseid häireid ja soole divertiikul- ehk sopististõbi.

Retseptidest on kõige rohkem välja kirjutatud retsepte toimeaine kategooriaga C, mis tähistab kardiovaskulaarsüsteemi ravimeid. Vaatlusajal ei ole 2%-le riskipatsientidest välja kirjutatud ühtegi retsepti. Vähemalt üks raviteenus on osutatud 87%-ile patsientidest. Riskipatsientide perearstid on 2017. aastal saanud keskmise

skoori perearstide kvaliteedi süsteemis 644 ning parim tulemus oli 772. Pooled perearstid on saanud tulemuseks rohkem kui 600 ning 25%-il perearstidest jääb skoor alla 184.

Tabel 4: Välja kirjutatud retseptide hinnad ja omaosalus patsiendi kohta (eurodes)

Tunnus	**Hospitaliseerimine aastal 2019**	**Min**	**Mediaan**	**Keskmine**	**Max**
Retseptide Hind Kokku	Ei	0	252,10	403,79	3 000
	Jah	0	339,24	527,07	3 000
Omaosalus Kokku	Ei	0	127,43	154,48	700
	Jah	0	163,19	189,95	700

Aastal 2018 on välja kirjutatud retseptide hind keskmiselt ühe patsiendi kohta 422,87 eurot ning keskmine omaosalus ühe patsiendi kohta 159,97 eurot. Tabelist 4 on näha, et hospitaliseeritud patsientidel on retseptide hind ja omaosalus olnud kõrgemad kui mittehopsitaliseeritutel. Samuti oli hospitaliseeritutel kõrgem mediaanhind ja -omaosalus. Välja ostmata retseptide osakaal oli 0,19 ning pooltel patsientidel jääb osakaal alla 0,16. Riskipatsientidest 15% jätab kõik välja kirjutatud retseptid välja ostmata.

3 Analüüs

Praktilise osa eesmärk on leida parim masinõppemeetod hospitaliseerimiste prognoosimiseks. Protsessi käigus katsetatakse ja võrreldakse erinevaid meetodeid tasakaalustamata andmete mudeldamiseks. Parima mudeli võimekust hinnatakse aastate 2020–2021 andmetel. Analüüsi osas kasutatatakse rakendustarkvara *R* ning närvivõrkude mudeldamiseks rakendustarkvara *Python*. Lisaks esimeses peatükis välja toodud pakettidele kasutatakse veel rakendustarkvara *R* pakette *tidyverse* ja *tidymodels* ning rakendustarkvara *Python* korral pakette *pandas* (McKinney ja Pandas Development Team, 2022) ja *Numpy* (Harris *et al.*, 2020).

3.1 Eeltöö

Praktilise osa käigus soovitakse leida masinõppemudel, mis eelneva aasta raviarvete andmete põhjal prognoosiks tõenäosust sattuda järgmisel aastal haiglasse. Mudeli treenimiseks jagatakse esmalt algne töödeldud andmestik treening- ja testandmeteks. Andmete kirjelduses tuli välja, et vaadeldavad andmed on tasakaalustamata ehk uuritava tunnuse klassid ei ole tasakaalus. Andmete jagamisel alamandmestikeks säilitatakse uuritava tunnuse jaotus. Treeningandmeteks võetakse algsetest andmetest 80% (78 173) ning testandmeteks jääb 20% andmetest (19 544). Andmestik jagatakse kaheks kasutades rakendustarkvara *R* paketti *rsample* funktsiooni *initial_split*. Alapeatükis 1.1 toodi välja, et testandmed on ainult mudeli võimekuse hindamiseks. Seega mudeli parameetrite hindamiseks jagatakse treeningandmestik veel kord kaheks: treeningandmestikuks, mis koosneb 62 538 vaatlusest, ning valideerimisandmestikuks, mis koosneb 15 635 vaatlusest.

Parima mudeli valimiseks tuleb leida sobilik kriteerium. Kasutatud mudelid hindavad vaatluse tõenäosust sattuda järgmisel aastal haiglasse. Tõenäosusi kasutatakse, et klassifitseerida patsiendid kõrge ja madala hospitaliseerimise riskiga patsientideks. Mudeldamiseks kasutatud andmed on tasakaalustamata, seetõttu ei ole sobi-

lik kasutada mudeli võimekuse hindamiseks õigsust. Lisaks on soov, et otsitav mudel leiaks võimalikult palju hospitaliseeritud inimesi ning samal ajal suudaks neid eristada mittehospitaliseeritutest. Mõõdiku valimisel tuleb lähtuda ka asjaolust, et hospitaliseeritud inimese klassifitseerimine mittehospitaliseerituks on kallim viga kui vastupidine klassifitseerimine. Vastavalt eelnevale leiti, et otsitakse mudelit lävendiga, mis annab minimaalse valepositiivsuse määra tingimusel, et valenegatiivsuse määr on väiksem või võrdne kui 0,05:

$$\min\{\text{Valepositiivsuse määr}|\text{Valenegatiivsuse määr} \leq 0,05\}. \tag{1}$$

3.2 Tulemused valideerimisandmetel

Mudeldamisel kasutatatakse tuntuid klassifitseerijaid, mis lisaks võimaldavad hinnata ka tõenäosust. Baasmudeliks on logistiline regressioon. Baasmudeli tulemust soovitakse paremaks saada kasutades täiend-log-log seosefunktsiooniga üldistatud lineaarset mudelit, otsustusmetsa ja kuni 3 peidetud kihiga pärilevi närvivõrke. Valideerimisandmestikku kasutatakse, et leida hüperparameetrite väärtused otsustusmetsa ja närvivõrkude korral. Lisaks leitakse sobivad lävendid kõikide meetodite puhul. Valideerimisandmestikku kasutatakse ka sobiva tasakaalustatud andmestiku valimiseks ning vaatluste kaalude valimisel.

3.2.1 Lävendimeetod

Esmalt leitakse tulemused kasutades lävendi meetodit (vt alapeatükk 1.8.1). Mudelid treenitakse treeningandmestikul. Sobiva lävendi leidmiseks ja mudelite hüperparameetrite hindamiseks kasutatakse valideerimisandmestikku.

Logistilise regressiooni ja **täiend-log-log** mudelite korral lisatakse mudelisse kõik andmestikus olevad tunnused. Antud mudelite korral valitakse valideerimisandmestikku kasutades ainult sobiv lävend. Lisaks proovitakse logistilise regressioo-

ni puhul parameetrite arvu kahandada lassoregressiooni abil. **Lassoregressiooni** korral kasutatakse andmestikku 190 tunnusega, kus kategoorilised tunnused on ümber defineeritud binaarseteks, ning otsitakse sobiv kahandav parameeter λ (vt alapeatükk 1.3). Sobivat λ väärtust otsitakse lõigust $[0{,}000001, 100]$. Lõik jagatakse 60 punktiks ning parim λ väärtus valitakse vastavalt mõõdikule (1). Selgub, et parima valepositiivsuse määra annab $\lambda = 0{,}002453751$, mille korral jääb mudelisse 46 tunnust (vt lisa 7).

Otsustusmetsa korral vaadeldakse tulemust vaikeväärtustega ning seejärel vaadeldakse, kas hüperparameetrite valimisega on võimalik tulemust paremaks muuta. Otsustusmets treenitakse rakendustarkvara *R* paketti *ranger* kasutades. Otsustusmetsa hüperparameetritest (vt alapeatükk 1.5) seadistatakse puude arv (*num.trees*), tunnuste arv (*mtry*), minimaalne lehe suurus (*min.node.size*) ja valimi suurus (*sample.fraction*). Mudeli treenimiseks kasutatakse parameetri *probability* puhul väärtust *TRUE*, et saada tõenäosushinnangud.

Tabel 5: Otsustusmetsa hüperparameetrite korral katsetatud väärtused, vaikeväärtused esile tõstetud

Hüperparameeter	**Väärtused**
Puude arv	**500**, 1 000, 2 000
Tunnuste arv	8, 10, **12**, 14, 16
Minimaalne lehe suurus	**10**, 20, 40, 80
Valikumäär	0,7; 0,8; 0,9; **1**

Tabelis 5 on kirjeldatud hüperparameetrite treenimiseks proovitud väärtused ning nendest on esile tõstetud mudeli vaikeväärtused. Treeningprotsessis vaadeldakse läbi kõik mudelid võimalike väärtuste kombinatsioonidega. Parima tulemuse annab otsustusmets parameetritega: *num.trees* = 500, *mtry* = 10, *min.node.size* = 40 ja *sample.fraction* = 0,7. Hüperparameetrite vaikeväärtustest (vt tabel 5) ja parima mudeli andnud väärtustest hakkab silma, et antud probleemi korral ei anna puude arvu suurendamine midagi juurde. Veel võib märgata, et minimaalne lehe suurus on 4 korda suurem kui vaikeväärtusega metsa puhul. Lisaks viitab suur minimaalne

lehe suurus sellele, et treenitud puud ei ole nii sügavad kui vaikeväärtustega metsa korral.

Närvivõrkude treenimisel tuleb valida mitmeid hüperparameetrite väärtuseid. Erinevalt otsustusmetsast puuduvad närvivõrkude mudeli parameetritele vaikeväärtused. Antud töö jaoks otsustatakse seadistada peidetud kihtide arv (mudeli sügavust), peidetud kihtide suurus (mudeli laiust), regulariseerimise parameeter ja miniploki suurus.

Tabel 6: Närvivõrkude hüperparameetrite katsetatud väärtused

Hüperparameeter	**Väärtused**
Peidetud kihtide arv	1, 2, 3
Peidetud kihi suurus	45, 95, 190, 380
Kantregulariseerija	0,00001; 0,0001; 0,001; 0,01
Miniploki suurus	32, 64, 128, 256

Antud töö raames vaadeldakse 1 kuni 3 peidetud kihiga närvivõrke. Peidetud kihi suuruse leidmiseks lähtutakse tunnuste arvust. Närvivõrkude treenimisel ei saa kasutada kategoorilisi tunnuseid, seega mudel treenitakse samal andmestikul kui lassoregressioon, mis sisaldab 190 tunnust. Kantregulariseerija väärtusena prooviti nelja väärtust (vt tabel 6). Miniploki suurusena vaadeldakse teoorias välja toodud väärtused (vt alapeatükk 1.6 ja tabel 6). Tabelis 6 on kirjeldatud võimalikud väärtused hüperparameetrite korral, kuid kõiki kombinatsioone treenimisel läbi ei proovita.

Treenimist alustatakse ühe peidetud kihiga pärilevi närvivõrkudest. Mudelite treenimisel võetakse epohhide arvuks 100. Kantregulariseerijat kasutatakse mudeldamisel nii, et esimesele ja teisele kihile lisati võrdsed regulariseerija väärtused. Parim tulemus saavutatakse miniploki suurusega 64, peidetud kihi suurusega 380 ning kantregulariseerijaga 0,00001. Kirjeldatud ülesehitusega närvivõrk andis valepositiivsuse määra **0**,**8631** ja valenegatiivsuse määra **0**,**0488**.

Kahe kihiga närvivõrkude korral jätkatakse eelmisel juhul parimaks osutunud mudeli treenimist. Miniploki suuruseks jääb 64, epohhide arvuks 100, esimese peidetud kihi suuruseks võetakse 380 ning esimese peidetud kihi regulariseerija väärtuseks 0,00001. Eesmärk on edasi arendada hetkel parimat tulemust. Teise peidetud kihi suuruseks katsetatakse taaskord väärtuseid tabelist 6. Teise ja kolmanda kihi regulariseerija väärtustena vaadeldi erinevaid kombinatsioone väärtustega tabelist 6. Selliselt tegutsedes saadakse parim tulemus mudeliga, mille teise kihi suurus on 45, teise kihi regulariseerija väärtus on 0,01 ning kolmanda kihi regulariseerija väärtus on 0,001. Kirjeldatud närvivõrk annab tulemuseks valepositiivsuse määra $\mathbf{0{,}8654}$ ja valenegatiivsuse määra $\mathbf{0,0480}$.

Kolme peidetud kihiga närvivõrgu puhul jätkatakse eelnevalt leitud parimat tulemust. Seekord fikseeritakse lisaks eelnevale teise kihi suurus 45 ja teise kihi regulariseerija väärtus 0,01 ning katsetatakse erinevaid väärtuseid kolmanda kihi suuruseks ja regulariseerijaks. Lisaks vaadeldakse viimase kihi erinevaid regulariseerija väärtuseid. Parima tulemuse annab närvivõrk peidetud kihtide suurustega: 380–45–45. Kihtidele lisatakse kantregulariseerijad: 0,00001–0,01–0,00001–0,0001. Kirjeldatud parameetritega saadakse tulemuseks valepositiivsuse määr $\mathbf{0,8602}$ ja valenegatiivsuse määr $\mathbf{0,0484}$. Seega kolme peidetud kihiga närvivõrk annab veidi parema tulemuse kui vaadeldud ühe peidetud kihiga närvivõrk. Antud töö raames rohkem kahe ja kolme peidetud kihiga närvivõrke ei uurita ning samuti ei vaadelda sügavamaid närvivõrke.

Tulemustest (vt tabel 7) selgub, et kõik meetodid annavad umbes sama valepositiivsuse määra. Kõige madalama tulemuse annab pärilevi närvivõrk 3 peidetud kihiga. Logistilise regressiooni puhul jääb silma, et tunnuste välja jätmine annab vaid veidi kehvema tulemuse kui kõikide tunnustega mudel. Otsustusmets vaikeväärtustega annab kõige kehvema tulemuse, mis jääb alla ka baasmudelile. Selgub, et otsustusmetsa puhul on oluline hüperparmeetrite seadistamine ning sellega parandatakse valepositiivsuse määra 0,01 võrra.

Tabel 7: Tulemused lävendi meetodiga

Meetod	**Valepositiivsuse määr**	**Valenegatiivsuse määr**	**Lävend**
Logistiline regressioon	0,8626	0,0484	0,081
Täiend-log-log	0,8624	0,0475	0,083
Logistiline regressioon 46 tunnusega	0,8669	0,0480	0,081
Otsustusmets vaikeväärtustega	0,8719	0, 0496	0,069
Otsustusmets seadistatud hüperparameetritega	0,8619	0,0484	0,075
Pärilevi närvivõrk 3 peidetud kihiga	0,8602	0,0484	0,096

3.2.2 Tasakaalustatud andmed

Järgmisena vaadeldakse mudelite tulemusi tasakaalustatud andmetel. Eesmärk on näha, kas andmete tasakaalustamine üle- ja alavalikuga (vt alapeatükk 1.8.2) aitab tulemusi tasakaalustamata andmetel paremaks muuta.

Andmete tasakaalustamiseks kasutatakse rakendustarkvara *R* paketti *groupdata2* funktsiooni *balance* (Olsen, 2021). Funktsiooni *balance* abil tekitatakse 3 andmestikku:

1. alavalikuga andmestik,
2. ülevalikuga andmestik,
3. ala- ja ülevalikuga andmestik.

Alavalikuga andmestiku korral jäetakse andmestikust välja enamusklassi vaatlused nii, et enamusklassi vaatluste arv oleks võrdne vähemusklassi vaatluste arvuga. Ülevalikuga tekitatud andmestiku korral paljundatakse vähemusklassi vaatluseid seni,

kuni vähemusklassis on sama palju vaatluseid kui enamusklassis. Ala- ja ülevalikuga andmestiku korral võetakse klasside vaatluste arvuks nende keskmine. Seejärel vähendatakse enamusklassi ja suurendatakse vähemusklassi vaatluste arvu, et saada tasakaalustatud andmestik.

Tabel 8: Tulemused tasakaalustatud andmetega

Andmed	**Meetod**	**Valepositiivsuse määr**	**Valenegatiivsuse määr**	**Lävend**
1	Logistiline regressioon	0,8621	0,0496	0,319
	Täiend-log-log	0,8664	0,0496	0,331
	Otsustusmets	0,8683	0,0488	0,324
2	Logistiline regressioon	0,8532	0,0492	0,322
	Täiend-log-log	0,8561	0,0495	0,334
	Otsustusmets	0,8663	0,0488	0,137
3	Logistiline regressioon	0,8630	0,0496	0,319
	Täiend-log-log	0,8667	0,0496	0,331
	Otsustusmets	0,8657	0,0484	0,196

Tasakaalustatud andmete korral vaadeldakse 3 meetodit: logistiline regressioon, täiend-log-log ja otsustusmets. Otsustusmetsa korral kasutatakse mudelit vaikeväärtustega. Mudeldamisel selgus, et lävend 0,5 ei anna piisavalt madalat valenegatiivsuse määra. Seetõttu leiti sobiv lävend, mis annab tulemuseks valenegatiivsuse määra väiksema kui 0,05. Tabelist 8 selgub, et parima tulemuse üldistatud lineaarsete mudelite korral annab ülevalikuga andmestik. Otsustusmetsa korral annab parima tulemuse ala- ja ülevalikuga andmestik. Samal ajal hakkab ka silma, et valepositiivsuse määra tulemus on umbes sama, mida nähti lävendi meetodi korral (vt tabel 7). Otsustusmetsa puhul võib näha, et tasakaalustatud andmete korral paraneb tulemus vaid veidi. Järelikult antud probleemi korral ei anna tasakaalustatud andmestik mudeldamisele midagi juurde. Tervel treeningandmestikul

treenitakse üldistatud lineaarsed mudelid ainult ülevaliku meetodil ja otsustusmets ala- ja ülevaliku meetodil.

3.2.3 Kaalutud vaatlused

Vaadeldud mudelitele saab lisada vaatlustele kaale, mida kasutatakse treeningprotsessis. Kaalude lisamise eesmärk on mudeli tähelepanu juhtida huvipakkuvale sündmusele treenimise ajal.

Antud töös katsetatakse vaid **otsustusmetsale** kaalude lisamist. Mudeldamisel kasutatakse hüperparameetrite korral vaikeväärtuseid. Mudelisse lisatakse parameeter *case.weights* ning hospitaliseeritutele katsetatakse kaalude väärtuseid $5, 10, 20, 40$. Mittehospitaliseeritutele jääb kaaluks 1. Peale mudeli treenimist oli vaja leida ka sobiv lävend, et valenegatiivsuse määr oleks väiksem kui 0,05.

Tabel 9: Otsustusmetsa tulemus kaalutud vaatlustega

Meetod	**Kaal**	**Valepositiivsuse määr**	**Valenegatiivsuse määr**	**Lävend**
Otsustusmets	5	0,8583	0,0492	0,211

Tulemusest selgub, et otsustusmetsa korral piisab hospitaliseeritutele kaaluks 5 (vt tabel 9). Taaskord saadakse valepositiivsuse määra tulemuseks ligikaudu 0,86. Kui võrrelda tulemust lävendi meetodil ja vaikeväärtustega treenitud otsustusmetsa tulemusega (vt tabel 7), siis on näha, et kaalude lisamine mudelisse annab veidi parema tulemuse. Kaalutud vaatlustega otsustusmets edestab veel ka tasakaalustatud andmetega otsustusmetsa tulemust (vt tabel 8).

3.3 Tulemused testandmetel

Järgnevalt võetakse vaatluse alla mudelid koos sobitatud hüperparameetritega, lävendite ja kaaludega ning treenitakse mudelid tervel treeningandmestikul. Seejärel

hinnatakse tulemused testandmetel. Testandmete tulemuste põhjal valitakse parim mudel hospitaliseerimise riski hindamiseks.

Tabel 10: Tulemused testandmetel

Meetod	**Lähenemine**	**Valepositiivsuse määr**	**Valenegatiivsuse määr**
Logistiline regressioon	Lävendi meetod	0,8622	0,0489
	Lassoregressioon	0,8669	0,0450
	Tasakaalustatud andmed	0,8577	0,0509
Täiend-log-log	Lävendi meetod	0,8627	0,0496
	Tasakaalustatud andmed	0,8606	0,0506
Otsustusmets	Lävendi meetod	0,8676	0,0479
	Tasakaalustatud andmed	0,8639	0,0456
	Kaalutud vaatlused	0,8612	0,0463
Närvivõrgud	Lävendi meetod	0,8587	0,0532

Tabelis 10 on välja toodud kõikide katsetatud meetodite tulemused testandmetel. Logistilise regressiooni puhul selgub, et vähendatud tunnustega mudel (lassoregressioon) saab kehvema valepositiivsuse määra kui kõikide tunnustega mudel lävendi meetodil. Selgub, et parima tulemuse annab logistiline regressioon tasakaalustatud andmetega. Täiend-log-log mudeli korral annab tasakaalustatud andmetega mudel parima tulemuse, kuid erinevus lävendi meetodist on vaid 0,021. Otsustusmetsa korral annab parima valepositiivsuse määra kaalutud vaatlustega mudel.

Tulemustest (vt tabel 10) hakkab silma, et tasakaalustatud andmetega mudelid edestavad lävendi meetodiga mudelite valepositiivsuse määrasid, kuid erinevused on väiksed. Otsustusmetsa korral on näha, et tasakaalustatud andmetega mudeli ja kaalutud vaatlustega mudelite korral on nii valepositiivsuse kui ka valenegatiiv-

suse määr väiksem kui lävendi meetodi korral. Närvivõrgu tulemusest on näha, et testandmetel on tulemus parem kui valideerimisandmete korral.

Tabelist 10 on näha, et tasakaalustatud andmed ning kaalutud vaatlused annavad hospitaliseeritute mudeldamisele juurde, kuid mitte märkimisväärselt. Seetõttu ei saa väita, et tasakaalustatud andmed või kaalutud vaatlused aitaksid paremini hospitaliseerituid eristada. Parima mudeli valimisel tuleb arvestada, et kokkuvõttes annavad kõik mudelid umbes sama tulemuse, seetõttu on mõistlik jääda ülesande baasmudeli ehk logistilise regressioon mudeli juurde. Eelnevat arvesse võttes valitakse parimaks mudeliks baasmudel ehk logistiline regressioon lävendi meetodiga.

3.4 Tulemused uutel andmetel

Järgmisena treenitakse valitud parim mudel tervel andmestikul, mis oli analüüsiks kasutada. Seejärel võetakse riskipatsientide 2020. aasta raviarvete andmed ja hospitaliseerimised aastast 2021 ning katsetatakse valitud parima mudeli võimekust uutel andmetel. Uute andmete korral on teada, et aastad 2020 ja 2021 võivad olla mõjutatud COVID-19 levikust.

Tabel 11: Tulemused uutel andmetel

		Prognoos		
		0	1	
Tegelik	0	11 801	67 991	79 792
	1	664	12 773	13 437
		12 465	80 764	93 229

Tabelis 11 on kirjeldatud tulemused uute andmete korral. Valepositiivsuse määraks saadakse **0,8521** ning valenegatiivsuse määraks saadakse **0,0494**. Tulemus on ligikaudu sama, mida nähti testandmete korral. Aasta 2020 andmete korral on valepositiivsuse määr isegi veidi madalam kui testandmete puhul. Tulemus annab lootust, et mudelit saab kasutada hilisemate aastate andmete korral. Lisaks võib järeldada, et COVID-19 levik ei oma mudeli võimekusele negatiivset mõju.

Lõpliku mudeli tulemustest uute andmete korral on näha, et perearstide jälgimise alla peaks minema 80 764 patsienti ehk 86,6% riskipatsientidest. Samal ajal väiksemat tähelepanu vajab 12 465 patsienti ehk 13,4% riskipatsientidest, kellest 5,3% vajab hospitaliseerimist järgmisel aastal. Tulemustest selgub, et mudeliga patsiente klassifitseerides saadakse võrdlemisi väike kokkuhoid töömahus. Seetõttu tuleb valminud mudelit kasutada pigem patsientide järjestamiseks riskiskoori alusel. Lisaks tuleb praeguste andmete juures leppida tähelepanuta jäävate haiglaravile sattujatega.

Kokkuvõte

Magistritöö eesmärk oli leida parim mudel hospitaliseerimiste prognoosimiseks. Esmalt tehti ülevaade masinõppe ning tuntumate klassifitseerimismeetodite teooriast. Klassifitseerijatest tutvustati lähemalt üldistatud lineaarseid mudeleid logit- ja täiend-log-log seosefunktsiooniga, otsustusmetsa ning närvivõrke. Lisaks tutvustati, mida peab jälgima mudeldades tasakaalustamata andmeid. Töö teises osas kirjeldati valimi moodustamise protsessi ning riskipatsiendi täpsemat definitsiooni. Veel tutvustati analüüsiks kasutatud andmestikku ning tehti ülevaade andmete töötlusest. Seejärel tehti kokkuvõte andmestikku sattunud riskipatsientidest.

Praktilises osas katsetati erinevaid meetodeid hospitaliseerimiste prognoosimiseks. Vaatluse alla võeti eelnevalt tutvustatud klassifitseerimismeetodid: logistiline regressioon, täiend-log-log mudel, otsustusmets ning närvivõrk. Klassifitseerimise puhul tuli arvestada, et hospitaliseeritud riskipatsiendi valesti kategoriseerimine on kallim viga kui mittehospitaliseeritu valesti kategoriseerimine. Lähtuvalt sellest fikseeriti enne mudeldamist, et otsitud mudel peab andma valenegatiivsuse määra alla 0,05 ning valepositiivsuse määra samal ajal võimalikult väikese. Teooria osas toodi tasakaalustamata andmete mudeldamiseks välja kaks meetodit: lävendi ja valikumeetod. Mõlema meetodi puhul katsetati logistilist regressiooni, täiend-log-log ning otsustusmetsa mudeleid. Lävendi meetodi korral vaadeldi veel lisaks logistilist regressiooni vähendatud tunnustega ja närvivõrke. Mudelite hüperparameetrid ja sobivad lävendid valiti valideerimisandmestikku kasutades ning seejärel hinnati mudelite võimekust testandmetel.

Tulemustest selgus, et kõik katsetatud mudelid ja meetodid andsid valepositiivsuse määra ligikaudu 0,86, kui valenegatiivsuse määr jääb alla 0,05. Kuna kõik tulemused on ligikaudu võrdsed, siis antud probleemi korral ei saa väita, et valikumeetodid tasakaalustamata andmete mudeldamisele midagi juurde annaks. Mudeldamise baasmeetodiks oli logistiline regressioon lävendi meetodiga. Baasmeetodi tulemust

ei suudetud katsetatud mudelitega märkimisväärselt parandada ning seetõttu valiti logistiline regressioon parimaks mudeliks.

Parim mudel treeniti tervel kasutuses olnud andmestikul ning seejärel leiti tema tulemus 2020. aasta andmetel. Uutel andmetel selgus, et valitud mudel saab 2021. aasta hospitaliseerimiste prognoosimisega sama hästi hakkama kui aastate 2018–2019 andmetel. Aasta 2021 tulemustest selgub, et perearstide tähelepanu vajab 86,6% riskipatsientidest. Mudeli poolt mittehospitaliseerituteks prognoositutest vajab hospitaliseerimist 5,3%. Tulemustest selgus, et mudelit on olulisem kasutada patsientide järjestamiseks. Valitud mudel võib tulevikus olla töövahend perearstidele, et leida patsiendid, kelle perearstlik jälgimine võib ära hoida tervise halvenemise haiglaravi vajaduseni.

Antud töö käigus vaadeldi vaid osa võimalikest mudelitest. Samuti katsetati kuni 3 peidetud kihiga närvivõrke. Edasi võiks uurida veel teisi võimalikke kuni 3 peidetud kihiga ning samuti ka sügavamaid närvivõrke. Lisaks võiks uurida klasside kaalumise mõju üldistatud lineaarsetele mudelitele ja närvivõrkudele. Autor leiab, et proovida võiks muuta ka andmestikku, näiteks lisada uusi kirjeldavaid tunnuseid. Lisaks võiks andmeid vaadelda aegreana.

Kasutatud materjalid

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu ja Xiaoqiang Zheng (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Data Science Estonia (2022). *Närvivõrkude ja masinõppe sõnastik*. URL: http://datasci.ee/masinoppe-sonastik/ (vaadatud 14.03.2022).

Dertat, Arden (2017). "Applied Deep Learning - Part 1: Artificial Neural Networks." URL: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (vaadatud 11.05.2022).

Eesti Haigekassa (2021). *Raviarvete ja lepingute andmevahetusteenused*. URL: https://www.haigekassa.ee/sites/default/files/RRL/2021/EHK_RTA_teenused_v4.52.pdf (vaadatud 10.05.2022).

– (2022a). *Eesti Haigekassa kindlustusliikide loetelu*. URL: https://www.haigekassa.ee/sites/default/files/kindlustusliigid.pdf (vaadatud 10.05.2022).

Eesti Haigekassa (2022b). *Perearsti kvaliteedisüsteem.* URL: https://www.haigekassa.ee/partnerile/raviasutusele/perearstile/perearsti-kvaliteedisusteem (vaadatud 09.05.2022).

– (2022c). *Tervishoid ja tervishoiuteenuste osutajad.* URL: https://www.haigekassa.ee/kontaktpunkt/arstiabi-valismaalasele-eestis/tervishoiususteemi-korraldus-eestis/tervishoid-ja (vaadatud 10.05.2022).

– (2022d). *Üldarstiabi rahastamise lepingud. Pearahasiseste tegevuste koodid.* URL: https://www.haigekassa.ee/partnerile/raviasutusele/perearstile/lepingud (vaadatud 10.05.2022).

Eesti Haigekassa ja Maailmapanga Grupp (2015). *Ravi terviklik käsitlus ja osapoolte koostöö Eesti tervishoiusüsteemis.* Kokkuvõttev aruanne. URL: https://www.haigekassa.ee/sites/default/files/Maailmapanga-uuring/veeb_est_summary_report_hk_2015.pdf.

Goodfellow, Ian, Yoshua Bengio ja Aaron Courville (2016). *Deep Learning.* http://www.deeplearningbook.org. MIT Press.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke ja Travis E. Oliphant (2020). "Array programming with NumPy". *Nature* 585.7825, lk. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hastie, T., R. Tibshirani ja J. Friedman (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Second Edition. Springer.

James, G., D. Witten, T. Hastie ja R. Tibshirani (2021). *An Introduction to Statistical Learning with Applications in R*. Second Edition. Springer. URL: https://www.statlearning.com.

Jong, P. de ja G. Z. Heller (2008). *Generalized Linear Models for Insurance Data*. Second Edition. Cambridge University Press.

Keras (2022). *EarlyStopping*. URL: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping (vaadatud 06.04.2022).

Kingma, Diederik P. ja Jimmy Lei Ba (2017). "Adam: A Method for Stochastic Optimization." URL: https://arxiv.org/abs/1412.6980 (vaadatud 09.05.2022).

Kuhn, Max ja Davis Vaughan (2022). "Logistic regression via glmnet." URL: https://parsnip.tidymodels.org/reference/details_logistic_reg_glmnet.html (vaadatud 16.05.2022).

Kuhn, Max ja Hadley Wickham (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. URL: https://www.tidymodels.org (vaadatud 06.04.2022).

Ling, Charles X. ja Victor S. Sheng (2010). "Cost-Sensitive Learning". Teoses: *Encyclopedia of Machine Learning*. Toim. Claude Sammut ja Geoffrey I. Webb. Boston, MA: Springer US, lk. 231–235. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_181. URL: https://doi.org/10.1007/978-0-387-30164-8_181.

Maa-amet (2021). *Geoportaal. Aadressid ja posti sihtnumbrid*. URL: https://geoportaal.maaamet.ee/est/Ruumiandmed/Aadressiandmed/Aadressid-ja-posti-sihtnumbrid-p582.html (vaadatud 20.10.2021).

Maailmapanga Grupp (2017). *Ravi juhtimine: suurenenud ravivajadusega patsientide ravi koordineerimine Eestis.* Eesti ravi juhtimise pilootprojekti 2017. aasta hindamisaruanne. URL: https://www.haigekassa.ee/sites/default/files/uuringud_aruanded/ECM-Pilot%20Evaluation_est_2018.pdf.

McKinney, Wes ja the Pandas Development Team (2022). "pandas: powerful Python data analysis toolkit". URL: https://pandas.pydata.org/docs/pandas.pdf (vaadatud 11.04.2022).

Milborrow, Stephen (2021). *Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'.* URL: https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf (vaadatud 06.04.2022).

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

Olsen, Ludvig Renbo (2021). *Creating Groups from Data.* URL: https://cran.r-project.org/web/packages/groupdata2/groupdata2.pdf (vaadatud 05.04.2022).

Probst, Philipp, Marvin Wright ja Anne-Laure Boulesteix (2019). "Hyperparameters and Tuning Strategies for Random Forest." URL: https://www.researchgate.net/publication/324438530_Hyperparameters_and_Tuning_Strategies_for_Random_Forest.

Provost, Foster (2000). "Machine Learning from Imbalanced Data Sets 101." URL: https://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf.

Python Software Foundation (2001-2022). *Python 3.9.7.* URL: https://www.python.org/.

R Core Team (2022a). *Documentation for package 'stats'*. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (vaadatud 06.04.2022).

– (2022b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/ (vaadatud 06.04.2022).

Ravimiamet (2022). *ATC puu*. URL: https://www.ravimiregister.ee/default.aspx?pv=Loendid.ATCPuu (vaadatud 09.05.2022).

Riigikantselei ja Justiitsministeerium (2022a). *Riigi Teataja. Eesti Haigekassa tervishoiuteenuste loetelu*. URL: https://www.riigiteataja.ee/akt/102042022001 (vaadatud 09.05.2022).

– (2022b). *Riigi Teataja. Haiglavõrgu arengukava*. URL: https://www.riigiteataja.ee/akt/13353001?leiaKehtiv (vaadatud 08.02.2022).

Silge, Julia, Fanny Chow, Max Kuhn ja Hadley Wickham (2021). *rsample: General Resampling Infrastructure*. URL: https://rsample.tidymodels.org/index.html (vaadatud 05.04.2022).

Statistikaamet (2022). *Statistika andmebaas. LES20. Vaesuse ja materiaalse ilmajäetuse määr elukoha järgi*. URL: http://andmebaas.stat.ee (vaadatud 08.02.2022).

Tervise ja Heaolu Infosüsteemide Keskus (2022). *NOMESCO kirurgiliste protseduuride klassifikatsioon*. URL: http://pub.e-tervis.ee/classifications/NCSP (vaadatud 10.05.2022).

Therneau, Terry, Beth Atkinson ja Brian Ripley (2022). *Recursive Partitioning and Regression Trees*. URL: https://cran.r-project.org/web/packages/rpart/rpart.pdf (vaadatud 06.04.2022).

Tiit, Ene-Margit ja Liina-Mai Tooding (2019). *Statistikaleksikon*. Tartu Ülikooli Kirjastus.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo ja Hiroaki Yutani (2019). "Welcome to the tidyverse". *Journal of Open Source Software* 4.43, lk. 1686. DOI: 10.21105/joss.01686.

Wright, Marvin N., Stefan Wager ja Philipp Probst (2021). *A Fast Implementation of Random Forests.* URL: https://cran.r-project.org/web/packages/ranger/ranger.pdf (vaadatud 06.04.2022).

Šteinmiller, Jekaterina (2021). "Riskipatsiendid tuleb kaasata oma ravi korraldusse." URL: https://www.haigekassa.ee/uudised/riskipatsiendid-tuleb-kaasata-oma-ravi-korraldusse (vaadatud 10.05.2022).

Lisa 1. Kaasuvad diagnoosid.

Järgnevas tabelis on välja toodud riskipatsientide valimi koostamisel ning hospitaliseerimise riski mudelites kasutatud diagnoosid ja nende täpsed määratlused raviarvetel esitatud RHK-10 koodide kaudu.

Tabel 12: Diagnoosid ja RHK-10 koodid

Diagnoosi nimetus	**RHK-10 kood**
Krooniline südamepuudulikkus	I11.0, I13.0, I13.2, I50.0, I50.1, I50.9
Astma	J45–J46
Ärevushäire	F40–F41
Peaaju transitoorse isheemia atakid ja -veresoonte haigused	G45, I60–69
Puriini- ja püramidiiniainevahetuse häired või podagra	E79, M10
Südame isheemiatõved	I20–I25
Artroosid	M15–M19
Kilpnäärme haigusseisund	E01–E05, E07, E06.1, E06.2, E06.3, E06.5, E06.9
Ateroskleroos	I65, I66, I70, I67.2, I73.9
Kodade virvendus ja laperdus	I48
Neuropaatiad	G50–G64
Aneemia	D50–D53, D55, D58, D61, D63, D64, D59.0, D59.1, D59.2, D59.4, D59.5, D59.6, D59.7, D59.8, D59.9, D60.0, D60.8, D60.9
Mao- ja söögitoru haigused	K21, K25.4, K25.5, K25.6, K25.7, K25.8, K25.9, K26.4–K26.9, K27.4–K27.9, K28.4–K28.9, K29.2–K29.9
Vertiigo ehk peapööritus	H81-H82, R42
Vähk[4] (Diagnoositud perioodil 01.01.2017–31.06.2018.)	C00–C97.99, D00–D09.99, D40–D49.99, D37–D39.99, Z51–Z51.99

[4]Valimi moodustamisel rohkem kui 7 teise diagnoosi arvutamisel ei arvestatud vähidiagnoosi.

Diagnoosi nimetus	**RHK-10 koodid**
Reumatoidartiit	M05–M06, M79.0
Osteoporoos	M80–M82
Südameklappide haigusseisundid	I34–I37
Alumiste hingamisteede kroonilised haigused	J40–J44, J47
Alajäsemete veenilaiendid	I83, I87.2
Südamehaigused	I44, I45, I47, I49
Koletsüstiit	K80, K81.1
Parkinsoni tõbi	G20–G22
Epilepsia	G40
Psoriaas	L40
Prostatiit	N40
Hemorroidid	I84
Maksahaigused	K70, K73–K74, K76, K71.3, K71.4, K71.5, K71.7, K72.1, K72.7, K72.9
Meeleoluhäired	F30–F39
Migreen	G43–G44
Neeru- ja ureeteri- e kusejuhakivi	N20
Inkontinentsus e kusepidamatus	R32, N39.3, N39.4
Rasvumus	E66
Somatoformsed häired	F45
Soole divertiikul- e sopististõbi	K57
Hüpotensioon	I95
Söömishäired	F50, R63.0
Kõne ja keele spetsiifilised arenguhäired	F80
Nägemise ja kuulmishäired	H54.1, H54.2, H54.0, H54.9, H90, H91
Dementsus	F00–F03, G30–G31, R54, F05.1
Funktsiooni nõrkus ja sellest tulenevad riskid	R54, W00, W04–W08, W10, W18, W19, R41.81, Z91.8
Alkoholi kuritarvitamine	F10, Z71.4, Z81.1
Ainete sõltuvus	F11-F19, F55, Z71.5, Z81.3, Z81.4

Lisa 2. Vaimsete häirete diagnoosid.

Järgnevas tabelis on välja toodud riskipatsientide valimi koostamisel kasutatud vaimsete häirete diagnoosid ja nende täpsed määratlused raviarvetel esitatud RHK-10 koodide kaudu.

Tabel 13: Vaimsete häirete diagnoosid ja RHK-10 koodid

Diagnoosi nimetus	**RHK-10 koodid**
Alkoholi tarvitamisest tingitud psüühika- ja käitumishäired	F10
Alkoholi kuritarvituse alane nõustamine ja järelevalve	Z71.4
Alkoholi kuritarvitus perekonnaanamneesis	Z81.1
Dementsus Alzheimeri tõvest	F00
Vaskulaarne dementsus	F01
Dementsus MK muudest haigustest	F02
Täpsustamata dementsus	F03
Alzheimeri tõbi	G30
Närvisüsteemi mujal klassifitseerimata muude degeneratiivhaigused	G31
Seniilsus e raukus	R54
Dementsusega deliirium	F05.1
Psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäired	F11–F19
Sõltuvust mittepõhjustavate ainete kuritarvitamine	F55
Ravimite kuritarvituse alane nõustamine ja järelevalve	Z71.5
Muude psühhoaktiivsete ainete kuritarvitus perekonnaanamneesis	Z81.3
Muude ainete kuritarvitus perekonnaanamneesis	Z81.4
Meeleoluhäired	F30–F39

Lisa 3. Kasutatud raviteenused.

Järgnevas tabelis on välja toodud hospitaliseerimise riski mudelites kasutatud raviteenuste koodid (Riigikantselei ja Justiitsministeerium, 2022a), sealhulgas ka raviteenustena kodeeritavad faktid patsiendi kohta, koos selgitustega. Raviteenuste hulgas vaadeldi ka üldarstiabi teenuseid (Eesti Haigekassa, 2022d), mis kuuluvad pearahasiseste tegevuste alla.

Tabel 14: Raviteenuste koodid koos selgitustega

Raviteenuse kood	**Selgitus**	**Tervishoiu-teenuse liik**
66100	Albumiin, valk	laboratoorsed uuringud
66101	Glükoos	
66102	Kreatiniin, uurea, kusihape	
66103	Bilirubiin, konjugeeritud bilirubiin	
66104	Kolesterool, triglütseriidid	
66105	Kolesterooli fraktsioonid: HDL, LDL	
66106	Ensüümid: ALP, ASAT, ALAT, LDH, CK, GGT, CK-MBa, alfa-amülaas	
66107	Naatrium, kaalium, kaltsium	
66108	Kloriid, liitium, laktaat, ammoonium	
66109	Raud, magneesium, fosfaat	
66110	Lipaas, pankrease amülaas	
66111	Antistreptolüsiin-O, reumatoidfaktor	
66112	C-reaktiivne valk	
66113	Happe-aluse tasakaal	
66114	Hemoglobiini derivaadid ja variandid: karboksühemoglobiin, methemoglobiin, fetaalne hemoglobiin	
66116	IgG uriinis või liikvoris	
66117	Albumiin uriinis (mikroalbumiin) või liikvoris	
66118	Glükohemoglobiin	

Raviteenuse kood	**Selgitus**	**Tervishoiu-teenuse liik**
66119	Immunofiksatsioon: liikvori oligoklonaalsed immunoglobuliinid, uriini või seerumi monoklonaalsed immunoglobuliinid	laboratoorsed uuringud
66120	Seerumi valkude elektroforees	
66121	Uriini või liikvori valkude elektroforees	
66122	Isoensüümide elektroforees	
66123	Spetsiifilised valgud 1: IgA, IgM, IgG, transferriin	
66124	Spetsiifilised valgud 2: tseruloplasmiin, haptoglobiin, C3, C4, tsüstatiin C, prealbumiin, alfa1-antitrüpsiin, immunoglobuliinide kapa- ja lambda-ahelad	
66125	Immunoglobuliinide alaklassid	
66126	Süsivesikdefitsiitne transferriin	
66127	Transferriini lahustuvad retseptorid	
66128	Angiotensiini muundav ensüüm	
66130	Hemoglobiin plasmas	
66131	Osmolaalsus	
66132	Krüoglobuliinid	
66133	Glükoos-6-fosfaatdehüdrogenaas	
66136	Porfüüriauuringud: delta-aminolevuliinhape, koproporfüriin	
66138	Ainevahetushaiguste sõeluuringud	
66139	Ainevahetushaiguste eriuuringud: aminohapped, suhkrud, puriinid ja pürimidiinid, orgaanilised happed, pika ahelaga rasvhapped, kreatiin, guanidinoatsetaat	
66200	Erütrotsüütide settekiiruse uuring	

Raviteenuse kood	**Selgitus**	**Tervishoiu-teenuse liik**
66201	Hemogramm (vere automaatuuring leukogrammita või kolmeosalise leukogrammiga)	laboratoorsed uuringud
66202	Hemogramm viieosalise leukogrammiga	
66207	Uriinianalüüs testribaga	
9040	Vere üldkolesterooli tase üle 5,0 mmol/l	perearstiarbi
9041	Mikroalbuminuuriatest positiivne	
9042	Määratud triglütseriidid	
9050	Glükohemoglobiin üle (või võrdne) 7,0%	
9101	Glükoosi analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9102	Kreatiniini analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9104	Üldkolesterooli analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9105	Kolesterooli fraktsioonide analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9117	Mikroalbuminuuria analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9118	Glükohemoglobiini analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	
9706	TSH analüüs, mis on mujal tervishoiuasutuses tehtud ja perearsti tervisekaardis dokumenteeritud	

Lisa 4. Kasutatud ATC-koodid.

Järgnevas tabelis on välja toodud hospitaliseerimise riski mudelites kasutatud ATC koodid, mida kasutati tunnuste loomisel, ja nende toimeaine rühm.

Tabel 15: ATC koodid

ATC koodid	**Rühma nimetus**
A01, A02, A03, A04, A05, A06, A07, A08, A09,A10, A11, A12, A14, A16	Seedekulgla ja ainevahetus
B01, B02, B03, B05	Veri ja vereloomeorganid
C01, C02, C03, C04, C05, C07, C08, C09, C10	Kardiovaskulaarsüsteem
D01, D02, D03, D04, D05, D06, D07, D08, D10, D11	Dermatoloogilised preparaadid
G01, G02, G03, G04	Urogenitaalsüsteem ja suguhormoonid
H01, H02, H03, H04	Süsteemsed hormoonpreparaadid v.a. suguhormoonid
J01, J02, J04, J05, J07	Infektsioonivastased ravimid süsteemseks kasutamiseks
L01, L02, L03, L04	Kasvajatevastased ja immunomoduleerivad ained
M01, M02, M03, M04, M05, M09	Skeleti- ja lihassüsteem
N01, N02, N03, N04, N05, N06, N07	Kesknärvisüsteem
P01, P02, P03	Parasiitide vastased ained
R01, R02, R03, R05, R06, R07	Hingamissüsteem
S01, S02	Meeleorganid
T01	Ravimisarnased preparaadid, testribad soodusega
V01, V03, V07	Varia

Lisa 5. Diskreetsete tunnuste teisendamine.

Järgnevas tabelis on välja toodud hospitaliseerimise riski mudelites kasutatud diskreetsed tunnused, mida töödeldi, koos neile rakendatud teisendusega.

Tabel 16: Diskreetsete tunnuste teisendamine

Tunnused	**Teisendus**
Operatsioon1kuni3h, *ArveidTeenuseTyyp20*, *sots_seisund*, *Raviteenus*:{66127, 9041, 9042, 9050, 9101, 9102, 9104, 9105, 9118, 9706}, *KirjutatudRetsepte*:{A02, A03, A06, A10, A11, B01, C02, C04, C05, D06, G03, G04, H03, J05, J07, M02, M03, M04, M05, N04, N07, R03, D, G, L, P}	0 (ei esine), 1 (esineb)
Operatsioone, *OperatsioonErakorraline*, *ArveidTeenuseTyyp16*, *Raviteenus*:{66110, 66111, 66113, 66117, 66123, 9040}, *KirjutatudRetsepte:*{A, B03, D01, H02, R06}	0,1,2+
ArveidTeenuseTyyp2, *EMOvisiit*, *Raviteenus:*{66100, 66108, 66109, 66118, 66200, } *KirjutatudRetsepteR01*	0,1,2,3+
ValtimatuAbiArveid, *Raviteenus*:{66103, 66207} *KirjutatudRetsepte:*{D07, N02, N03, N06, }	0,1,2,3,4+
Raviteenus:{66105, 66201} *KirjutatudRetsepteJ01*	0,1,2,3,4,5+
Raviteenus66101	0,1,2,3,4,5,6+
Raviteenus:{66112, 66202}, *KirjutatudRetsepte:*{C01, C03, C07, C08, C10, M01, S01}	0,1,2,...,7+
KirjutatudRetsepteN05	0,1,2,...,8+
Raviteenus66107	0,1,2,...,10+
KirjutatudRetsepteC09, *Raviteenus*:{66102, 66106}	0,1,2,...,11+
AmbArveidEriarst	0,1,2,...,12+
AmbArveidPerearst	0,1,2,...,16+

Lisa 6. Kategooriliste tunnuste moodustamine.

Järgnevas tabelis on välja toodud hospitaliseerimise riski mudelites kasutatud tunnused, mis teisendati kategooriliseks.

Tabel 17: Tunnuste teisendamine kategooriliseks

Tunnused	**Teisendus**
KirjutatudRetsepteC08	0, [1,3], [4,6], [7,∞]
KirjutatudRetsepteC01, *KirjutatudRetsepteC03*, *KirjutatudRetsepteC07*, *KirjutatudRetsepteC09*, *KirjutatudRetsepteC10*	0, [1,3], [4,6], [7,9], [10,∞)
KirjutatudRetsepteS01	0, [1,6], [7,12], [13,∞)
Raviteenus66107	0, [1,2],[3,4], [5,6], [7,8], [9,10], [11,12], [13,∞)
Voodipaevi	0, (0,30], (30,∞)
VoodipaeviMuu	0, (0,7], (7,∞)
VoodipaeviIntensiivravi	0, (0,1], (1,∞)
arve_ VPT	0, (0,10], (10,∞)
arve_ VP_ OA	0, (0,15], (15,30], (30,∞)

Lisa 7. Lassoregressiooniga alles jäänud tunnused.

Järgnevalt on välja toodud hospitaliseerimise riski mudelites kasutatud tunnuste nimetused, mis jäid lassoregressiooniga mudelisse.

Tabel 18: Lassoregressiooniga allesjäänud tunnused

Sugu	*Vanus*	*VältimatuAbiArveid*
ArveidTeenuseTyyp2	*ArveidTeenuseTyyp20*	*AmbArveidPerearst*
AmbArveidEriarst	*Raviteenus66200*	*Raviteenus66106*
Raviteenus66112	*Raviteenus66207*	*Raviteenus66111*
Raviteenus9041	*DiagnoosiGrArv*	*Diagnoos südameisheemia*
Diagnoos artroosid	*Diagnoos ateroskleroos*	*Diagnoos kodade virvendus ja laperdus*
Diagnoos südameklappide haigusseisundid	*Diagnoos alumiste hingamisteede kr haigused*	*Diagnoos alajäsemete veenilaiendid*
Diagnoos koletsüstiit	*Diagnoos maksahaigused*	*Diagnoos dementsus*
RetseptideHindKokku	*OmaosalusKokku*	*KirjutatudRetseptеA02*
KirjutatudRetseptеN05	*KirjutatudRetseptеB01*	*KirjutatudRetseptеH02*
KirjutatudRetseptеA10	*KirjutatudRetseptеM01*	*KirjutatudRetseptеN02*
KirjutatudRetseptеN03	*KirjutatudRetseptеN06*	*KirjutatudRetseptеC02*
Kindlustuse Liik Töövõimetu	*Vaesus*	*ValjaOstmata*
Voodipäevi (0.30]	*VoodipäeviMuu* $(7.Inf]$	*arve_VP_OA_0*
KirjutatudRetseptеC01 (3.6]	*KirjutatudRetseptеC09* (0.3]	*KirjutatudRetseptеC03* $(6, 9]$
Raviteenus66107 $(2, 4]$		

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kadi Kilgi,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Hospitaliseerimise riski prognoosimine krooniliste haigustega patsientidel", mille juhendajad on Mark Gimbutas ja Krista Fischer, 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.

Kadi Kilgi
17.05.2022