TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

Silvia Pihu

Gestatsioondiabeedi ja makrosoomia prognoosimine ning nende riskitegurite analüüs masinõppe meetoditega

Magistritöö (15 EAP)

Juhendajad:

PhD Sven Laur (TÜ ATI),

PhD dr Kristiina Rull (TÜ Kliinikum)

Tartu 2020

Gestatsioondiabeedi ja makrosoomia prognoosimine ning nende riskitegurite analüüs masinõppe meetoditega

Lühikokkuvõte:

Makrosoomiat ehk liiga suurt vastsündinut käsitletakse töös gestatsioonilise vanuse suhtes. See põhjustab sageli probleeme sünnitusele, nii emale kui lapsele. Seetõttu on vajalik makrosoomse lapse sündi (varakult) ette ennustada, et kasutada dieetravi või ravida medikamentide abil makrosoomia riskiteguriks olevat gestatsioondiabeeti (GDM; rasedusdiabeet) või vähemalt plaanida keisrilõiget. Kasutades teadaolevaid ja oletatavaid riskifaktoreid prognoositi GDMi ja makrosoomiat erinevate masinõppe mudelite abil Eestis kogutud raseduste andmetel (2012-2018, 4787 raseduse andmed). Kõige parema ennustusvõimega oli kasutatutest juhusliku metsa meetod, millega ROC-kõvera aluseks pindalaks saadi GDM jaoks 0,96 ja makrosoomia jaoks 0,92. Kõige olulisemateks tunnusteks ja seega kõige olulisemateks riskifaktoriteks makrosoomse lapse sünniks osutusid GDM ja selle korrektne diagnoosimine, ema kehamassiindeks enne rasedust, ema vanus, varasem makrosoomse lapse sünd ja paastusuhkur raseduse alguses.

Võtmesõnad:

Gestatsioondiabeet, makrosoomia, masinõpe, andmekaeve, binaarne klassifikatsioon, tunnuste valik

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika; B570 Sünnitusabi, günekoloogia, androloogia, paljunemine, seksuaalsus

Prediction of Large-for-Gestational-Age and Gestational Diabetes Mellitus with Machine Learning Methods and Analysis of Risk Factors

Abstract:

Large-for-gestational-age (LGA) may cause problems for both baby and mother during delivery, therefore the best solution is to predict and avoid it (by diet, cure of GDM, etc.) or at least use planned Caesarian section. Gestational diabetes (GDM) is known as a major risk factor for too large baby. Different machine learning algorithms were used to predict GDM and LGA on Estonian pregnancies and newborn data from 2012 to 2018 (4787 cases), using their risk factors. The best results were obtained by random forest method (AUC for GDM 0.96 and for LGA 0,92). The major risk factors for LGA occurred to be GDM and its correct diagnosing, the body mass index of the mother (before pregnancy), having large baby in previous pregnancy, the age of mother and the blood sugar level registered at the beginning of pregnancy.

Keywords: large for gestational age, macrosomia, gestational diabetes mellitus, machine learning, data mining, binary classification, feature selection

CERCS: P160 Statistics, operation research, programming, actuarial mathematics; B570 Obstetrics, gynaecology, andrology, reproduction, sexuality

Sisukord

1. Sissejuhatus

Makrosoomseks võib nimetada loodet või vastsündinut, kes on liiga suur ja võib põhjustada probleeme sünnitusele nii emale kui lapsele. Gestatsioondiabeet on diabeet, mis tekib raseduse ajal ja see on üheks põhiliseks makrosoomse lapse sünni riskiteguriks.

Töö eesmärgiks on ennustada gestatsioondiabeedi ja makrosoomia teket riskitegurite alusel, analüüsida makrosoomia ja gestatsioon seost ning hinnata riskitegurite olulisust kasutades erinevaid andmeteaduse ja masinõppe meetodeid.

Tööle on lisatud sõnastik nii olulisemate rasedus- ja sünnitusalaste kui ka masinõppe ja andmeteaduse alaste terminitega.

Teoreetilise tausta peatükis käsitletakse masinõppe olemust, binaarse klassifikatsiooni algoritme, mudelite hindamiskriteeriume, aga samuti ka makrosoomia ja gestatsioondiabeedi olemust ja varasemaid uuringuid.

Materjali ja metoodika osas kirjeldatakse andmestikku, tunnuseid ja andmeanalüüsi meetodeid. Tulemustes tuuakse välja gestatsioondiabeedi ja makrosoomia seosed riskiteguritega, prognoositakse gestatsioondiabeeti ja makrosoomiat masinõppe algoritmide abil ja analüüsitakse riskitegurite rolli.

Järelduste ja arutelu osas võetakse eelnev kokku ja tuuakse välja olulisemad järeldused.

2. Mõisted, terminid ja kasutatavad lühendid

ACC – vt õigsus

Asfüksia - loote /vastsündinu verevoolu- või gaasivahetuse häire kas vahetult enne sündi, sünni ajal või pärast sündi, mille tõttu võivad tekkida tõsised süsteemsed või neuroloogilised tagajärjed [1].

AUC – vt ROC kõvera alune pindala

Bonferroni korrektsioon – mitmese keskmiste võrdlemise puhul, kui ühe võrdluse puhul etteantud olulisuse tõenäosuse lävend $\alpha=0,05$, siis see kehtib üksikvõrdlusele, kuid mitte tunnuste komplektile ühel obkektil, seega n võrdluse korral tuleb üksikvõrdlustel võtta olulisuse nivooks α/n [2].

DT – vt otsustuspuu

Fisheri täpne test - $\chi 2$-testi analoog, permutatsioonitest, kus arvutatakse kõik võimalikud kahemõõtmelised sagedustabelid ning rea- ja veerusageduste tõenäosused nullhüpoteesi eeldusel [3].

GDM – vt gestatsioondiabeet

Gestatsioondiabeet (GDM) – raseduse ajal tekkiv süsivesikute ainevahetushäire, mis taandub pärast sünnitust [4]. Vt ka lk 17

Gestatsiooniline vanus (GA) - raseduse kestus (enamasti mõõdetuna päevades), alates viimase menstruatsiooni esimesest päevast kuni sünnituseni [5].

Glükoosi taluvuse test (GTT) - test gestatsioondiabeedi avastamiseks, järjestikuselt määratakse tühja kõhu veresuhkur ja veresuhkru väärtused kindlate ajavahemike järel pärast glükoosi tarvitamist [6].

GNB – Gaussi naiivne Bayes, naiivse Bayesi meetod seotuna Gaussi jaotusega, vt lk 10

GTT – vt glükoosi taluvuse test

HP – Happy Pregnancy projekti raames kogutud andmestiku osa, vt lk 20

Hüpoglükeemia - veresuhkru langemine alla normi [7].

Juhuslik mets (RF) – käesolevas töös juhuslike otsustuspuude kogum, vt ka lk 11

Kalibreerimine (mudelil) – mudeli seadistamine vastavalt prognooside tõenäosustele, vt lk 14

Kehamassiindeks (KMI) - enim kasutatud kehakaalu adekvaatsust hindav näitaja, mille saamiseks jagatakse kehamass kilogrammides pikkuse ruutväärtusega meetrites [8].

KMI – vt kehamassiindeks

KNN – vt lähimate naabrite meetod

Korrelatsioon - juhuslike suuruste vahel esinev statistiline seos [9].

LDA – vt lineaarne diskriminantanalüüs

Lineaarne diskriminantanalüüs (LDA) – lineaarne klassifikatsioonimeetod, vt lk 12

Logistiline regression (LR) – binaarse klassifikatsiooni meetod, prognoosib uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt faktorite väärtuse muutumisest, vt ka lk 12

LR – vt logistiline regressioon

Lähimate naabrite meetod (KNN) - klassifikatsioonimeetod, mis otsib mitmemõõtmelises tunnusruumis lähimaid naabreid, vt lk 10

Makrosoomia, makrosoomne laps – käesolevas töös vastsündinu sünnikaaluga, mis langeb vastavalt gestatsioonilisele vanusele ja kasvukõveratele 0.97 kvantiili vastavalt Euroopa laste kasvukõveratele [10]. Vt ka lk 16

Naiivne Bayes (NB) – tõenäosustel põhinev klassifikatsioonimeetod, vt lk 10

NB – vt naiivne Bayes

Otsustuspuu (DT) – klassifikatsioonimeetod, mis on üles ehitatud puukujuliselt küsimuste-vastustena, vt lk 11

Paastusuhkur – käesolevas töös veresuhkru näit, mida määratakse tühja kõhuga

PCOS – vt polütsüstiliste munasarjade sündroom

Perineum e lahkliha – vaagnapõhja pehmed koed, mis paiknevad väliste suguelundite ja päraku vahel [11].

Polühüdramnion – liigne lootevee kogus, mis võib põhjustada probleeme sünnitusel ja isegi loote surma [12].

Polütsüstiliste munasarjade sündroom (PCOS) - hormonaalne häire, millele on iseloomulikud suurenenud meessuguhormoonide sisaldus, tsüstidega munasarjad ja ovulatsiooni puudumine. Sellega võib kaasneda diabeet ja rasvumine. [13]

Preeklampsia – rasedusaegne veresoonkonna patoloogia, mille puhul tekib veresoonte ahenemine ja trombotsüütide agregatsioon, mis halvendab oluliselt platsenta verevarustust ning pidurdab sellega loote kasvu ja hapnikuga varustamist. Preeklampsia avaldub enamasti 32.–36. rasedusnädalal tursete, vererõhu tõusuna ja valguna uriinis. Preeklampsia tagajärjel võib tekkida ema ja loote elu ohustav eklampsia (krambihoog).[5]

Regressioonanalüüs - mingi nähtuse kirjeldamiseks vaatlusandmete põhjal mudeli matemaatilise kuju ning mudeli parameetrite leidmine. Mudelil on deterministlik komponent ehk see, mis meid huvitab, ja juhuslik komponent. Lineaarse regressioonmudeli üldkuju on:

$$y = \alpha x + \beta + \varepsilon,$$

kus α ja β on mudeli parameetrid ning ε juhuslik liige. [9]

RF – vt juhuslik mets

Ristvalidatsioon – masinõppe meetod, vt lk 14

ROC-kõvera alune pindala (AUC) – klassifikatsioonimudeli hindamiskriteerium, vt lk 13

Saagis – vt tundlikkus

Spetsiifilisus (TNR) – klassifikatsioonimudeli hindamiskriteerium, vt lk 13

Studenti t-test – võrdleb kahe valimi keskmisi ja hindab selle järgi, kas üldkogumite keskmised on võrdsed või mitte [9].

SVM –vt tugivektor-masin

Šansside suhe (OR, *odds ratio*) - näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss faktoritele eksponeerituil võrreldes mitteeksponeeritutega. Sündmuse toimumise šanss näitab, mitmel juhul sündmus toimub võrreldes sellega, mitmel juhul ta ei toimu. [3]

Z-väärtus – näitab parameetri väärtuse kaugust keskmisest standardhälbe ühikutes [14]. Käesolevas töös on sünnikaalu puhul on kasutatud kaugust keskmisest vastavalt gestatsioonilisele vanusele, mitte üldkeskmisest.

Testandmed – masinõppes andmed, mis eraldatakse enne mudeli treenimist ja millel pärast mudeli treenimist hinnatakse selle ennustusvõimet

Treeningandmed - masinõppes andmed, millel mudelit treenitakse (mudel õpib)

t-test – vt Studenti t-test

Tugivektor-masin (SVM) - tugivektoritel põhinev klassifikatsioonimeetod, vt lk 11

Tundlikkus (TPR) – klassifikatsioonimudeli hindamiskriteerium, vt lk 13

Täpsus (PPV) – klassifikatsioonimudeli hindamiskriteerium, vt lk 12

Valenegatiivsete määr (FNR) – klassifikatsioonimudeli hindamiskriteerium, vt lk 13

Valepositiivsete määr (FPR) – klassifikatsioonimudeli hindamiskriteerium, vt lk 13

Võrkotsing (*grid search)* – otsing leidmaks parimaid mudeli (hüper)parameetreid, kus katsetatakse andmetel läbi hulk erinevaid mudeli parameetreid ja otsitakse kombinatsioon, kus ennustusvõime on parim, tavaliselt kasutades ristvalidatsiooni

Õigsus (ACC) – klassifikatsioonimudeli hindamiskriteerium, vt lk 12

Õlgade düstookia - sünnituse peatumine, mille põhjuseks on loote õlgade peetumine ema häbemeliiduse taga [11].

χ^2-test - kahemõõtmelise sagedustabeli alusel tehtav test tunnuste seotuse statistilise olulisuse määramiseks [3].

3. Teoreetiline taust

3.1. Andmeteadus ja masinõpe

3.1.1. Mõisted

Andmeteadus (*data science)* on lai mõiste – see hõlmab kõiki tegevusi, mis aitavad andmete põhjal kasulikke otsuseid teha, et seatud eesmärke saavutada [15].

Masinõpe (ML, *machine learning*) põhineb arvutite õppimisvõimel, olles algoritmiliste meetodite kogum raskelt defineeritavate lahendustega probleemide jaoks. Tom Mitchell on 2006. a sõnastanud masinõppe eesmärgi: "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" [16] lk 2. Peter Flach ütleb: „Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience." [17] lk 12.

Üks osa masinõppest on seotud andmeteadusega ja siin on eesmärk sellise algoritmi loomine, mis suudaks andmetest leida olulise info, nö signaali ja selle põhjal midagi ennustada. Andmeanalüüsi, mis kasutab masinõppe meetodeid, nimetatakse sageli ka andmekaeveks (*data mining*). Aluseks võetakse tavaliselt mingi eelnevalt määratletud parameetritega mudel ja arvuti õpib (treening)andmete põhjal, optimeerides parameetreid ja tulemuslikkuse näitajaid [18]. Enamasti katsetatakse tulemust testandmetel. Õppimine võib laias laastus defineerituna olla juhendatud, juhendamata või stiimulõpe [19]. Kasutusel olevate erinevate algoritmide hulk on väga suur.

Juhendatud õppe meetodid on kõige laiemalt levinud. Nende hulka kuuluvad näiteks spämmifiltrid, näotuvastussüsteemid, aga ka meditsiiniline diagnoosimine [20]. Põhimõtteliselt koosnevad treeningandmed x ja y paaridest ja eesmärk on ennustada y tunnuseid x-i põhjal. Nii x kui ka y võivad seejuures olla nii vektorid, maatrikid kui ka keerukamad objektid nagu dokumendid, pildid, DNA-järjestused vm [20]. Üks osa juhendatud õppest hõlmab ka närvivõrke ja süvaõpet.

Juhendamata õpe otsib andmetes ilma „õigeid vastuseid" ette andmata mustreid või struktuure. Sellesse kuuluvad näiteks klaster- ja peakomponentanalüüs.

Stimuleeritud õpe on teatud mõttes vahepealne, treeningandmetes puuduvad nö õiged vastused, need annavad ainult vihje, kas õppetegevus on olnud korrektne või mitte [20]. See on eelmistest vähemlevinud ja tegelikult võib seda interpreteerida kas juhendatud või juhendamata õppena.

Niisiis, masinõppe olemus andmeteaduses seisneb vastavalt eesmärgile andmete/tunnuste põhjal mudeli koostamises, mis võimaldaks midagi ennustada. Et seda teha, on meil vaja õppimise algoritmi. [17]

3.1.2. Klassifikatsiooni algoritmid

3.1.2.1. Binaarse klassifikatsiooni mudelid

Sageli on vaja midagi klassifitseerida kaheselt ehk binaarselt: nt kas e-kiri on spämm või ei ole, meditsiinis kas on haigus või ei ole. Ehk siis sihttunnus on binaarne.

Sellisel juhul enimkasutatavad algoritmid on järgmised:

Lähimate naabrite meetod (*K Nearest Neighbors*, KNN)

Kasutatakse nii klasterdamisel kui ka klassifitseerimisel. See on tegelikult mudelivaba meetod, kus lähtutakse lihtsalt kaugusest mitmemõõtmelises tunnusruumis ja selle puhul määratakse testobjekti kuuluvus nii, et arvutatakse kaugused kõikidest treeningobjektidest, leitakse k (etteantud arv) lähimaid naabreid nende hulgas ja siis kasutatakse enamuse reeglit klassikuuluvuse määramiseks vastavalt treeningandmete klassikuuluvusele [21].

Naiivne Bayes'i meetod (*Naïve Bayes*, NB)

Hästi tuntud Bayes'i teoreem käsitleb sündmuse tõenäosust eeldusel, et sündmus on seotud juba toimunud sündmustega/faktoritega:

$$\boldsymbol{P(A|B) = P(B|A) * \frac{P(A)}{P(B)}}$$

kus P on tõenäosus, A on uuritav tunnus või sündmus ja B on juhtunud sündmus, sõltumatu tunnus, võib ka olla vektor, pilt vm. Ehk siis leiame A tõenäosuse eeldusel, et B on toimunud. Info vaatlusandmetest kombineeritakse eelinfoga (*apriori* infoga) sõltumatute tunnuste kohta. Kui

sõltumatuid tunnuseid on rohkem kui üks, siis selle meetodi puhul eeldatakse (naiivselt, sellest ka nimetus), et need on omavahel sõltumatud ja võrdse mõjuga hüpoteesile, uuritavale tunnusele [22], [23]. **Gaussi naiivse Bayesi meetod** (GNB) lihtsustab meetodit nii, et kasutatakse Gaussi jaotust, et minimeerida ruutvigade summat [23].

Otsustuspuu meetod (*Decision Tree*, DT)

Otsustuspuu on traditsiooniline masinõppe algoritm, kus aluseks on tipp, mis jaguneb harudeks, mis omakorda lõpevad tippudega, millest igaüks võib edasi haruneda või lõppeda. Iga tipu juures on küsimus, mille vastus näitab, millist haru tuleb edasi minna kuni otsuseni. Otsustuspuu ei ole sageli väga hea täpsusega, kuid on kompromiss täpsuse ja arusaadavuse vahel. [24],[25]

Juhusliku metsa meetod (*Random Forest*, RF)

Mitme otsustuspuu kombineerimisel saame otsustuspuude metsa. Tegelikult võib juhuslikku metsa kasutada ka mõnede teiste mudelite, nt regressioonimudelite kombineerimiseks. Mudelite kombineerimist üldisemalt nimetatakse ansamblimeetoditeks, ansamblimeetodeid on lisaks juhuslikule metsale ka teisi.

Juhusliku metsa puhul kasutatakse juhuslikku valikut nii vaatluste kui tunnuste puhul. Sellega on võimalik saavutada suurem sõltumatus andmetes sisalduvast mürast ja ekstreemsetest vaatlustest ning liigsest sobitamisest andmetele. Ka tuleb juhuslik mets paremini toime tasakaalustamata andmetega. Juhuslikkuse tõttu on iga puu sõltumatu mudel ja nende põhjal koostatud koondmudel on treeningandmetest vähem sõltuv. [24],[25]

Tugivektor-masina meetod (*Support Vector Machine*, SVM)

Tugivektor-masin on defineeritud nn tuuma (*kernel*) poolt, mis projekteerib andmed kõrgedimensioonilisse vektorruumi. See on meetod, mis otsib tugivektoreid, nimelt klasside äärealadel asuvaid vaatlusi, mille abil saaks klasse eristada. Oluline on otsida suurimat vahet klasside vahel. Tugivektor-masinate eeliseks on stabiliseerimine juhtudel kui probleem on mitmemõõtmeline ja lineaarne klassifikatsioon kaldub ülesobitusele. [24],[25]

Logistiline regressioon (*Logistic Regression*, LR)

See mudel prognoosib uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt pideva(te) faktori(te) väärtuse muutumisest. Sarnaneb (lineaarse) regressiooniga, kuid sihttunnus on binaarne, mitte pidev. [25],[3]

Lineaarne diskriminantanalüüs (*Linear Discrimant Analysis*, LDA)

Mudel leiab lineaarse kombinatsiooni parameetritest/faktoritest, nn diskrimineeriva funktsiooni, mis eristab kahte või enamat klassi [26]. Kasutatakse nii klassikalises statistilises analüüsis kui ka masinõppes.

3.1.2.2. Klassifikatsiooni hindamiskriteeriumid

Klassifikatsiooni hindamise aluseks on valdavalt nn segadusmaatriks (*confusion matrix*), mis näeb välja järgmine:

		Tegelik	
		Negatiivne	Positiivne
Prognoositud	Negatiivne	TN (*True negative*) – õige negatiivne	FN (*False negative*) – vale negatiivne
	Positiivne	FP (*False positiive)* – vale positiivne	TP (*True positive*) - õige positiivne

Sellest tuleneb ka enamik teisi kriteeriume (refereeritud peamiselt [27] järgi).

Õigsus (*accuracy*, ACC) on õigete määrangute proportsioon juhtude koguarvust:

$$ACC = \frac{TN + TP}{N}$$

kus N on uuritud juhtude koguarv ehk TN+FN+FP+TP

Täpsus (*precision, positive predictive value*, PPV) on tegelike positiivsete proportsioon prognoositud positiivsete hulgas:

$$PPV = \frac{TP}{TP + FP}$$

Tundlikkus ehk saagis, ühtlasi ka õigete positiivsete määr (*sensitivity*, *recall*, TPR) on õigesti prognoositud positiivsete proportsioon tegelike positiivsete hulgas ehk meditsiini mõistes „ülesleitud“ haiged kõikide haigete hulgas:

$$TPR = \frac{TP}{TP + FN}$$

Spetsiifilisus, ühtlasi ka õigete negatiivsete määr (*specificity*, TNR) on õigesti prognoositud negatiivsete määr tegelike negatiivsete hulgas:

$$TNR = \frac{TN}{TN + FP}$$

Valepositiivsete määr (FPR) on valesti prognoositud positiivsete proportsioon tegelike negatiivsete hulgas ehk nö valehäire määr:

$$FPR = \frac{FP}{FP + TN}$$

Valenegatiivsete määr (FNR) on valesti prognoositud negatiivsete proportsioon tegelike positiivsete hulgas, seega avastamata (haigus)juhtumite määr:

$$FNR = \frac{FN}{FN + TP}$$

F1 on täpsuse ja tundlikkuse harmooniline keskmine:

$$F1 = 2 * PPV * \frac{TPR}{PPV + TPR}$$

ROC-kõvera alune pindala (AUC) on tundlikkuse ja spetsiifilisuse suhet iseloomustava kõvera alune pindala.

3.1.2.3. Ristvalidatsioon ja mudeli kalibreerimine

Masinõppe mudelite kasutamise puhul klassifitseerimisel võib tekkida nii alasobitus kui ka ülesobitus. Alasobitus tähendab seda, et algoritm ei suuda leida andmetes leiduvat mustrit ehk mudel on liiga lihtne või on liiga palju puuduvaid andmeid või erindeid (*outliers*), selle vastu aitab tegelemine puuduvate andmetega ja erinditega, lineaarse mudeli asendamine keerulisemaga, polünomiaalsete tunnuste kasutamine jne.

Ülesobitus tähendab seevastu, et mudel on liiga keeruline ja kajastab liiga palju andmetes esinevat müra. Tihti avaldub see selles, et treeningandmete hindamiskriteeriumid (õigsus jm) on väga head, kuid testandmetel märksa kehvemad.

Ristvalidatsioon on üks viise ülesobitusega tegelemiseks. Treeningandmed jagatakse mitmeks osaks ning üks osa neist eraldatakse ajutisteks testandmeteks ja niimoodi tehakse läbi kõik kombinatsioonid (olenevalt osade arvust) [28].

Kalibreerimine on protsess, kus võetakse juba treenitud mudel ja järeltöödeldakse seda täpsustamaks prognoosimise tõenäosust. Ehk siis ideaaljuhul oleks perfektne lineaarne sõltuvus tõenäosuse ja positiivsete osakaalu vahel [29]. Teostamiseks tuleb prognoosid grupeerida vastavalt nende prognoosi tõenäosustele. Seejärel arvutatakse positiivsete osakaal grupis ja lõpuks gruppi kuuluvate juhtude keskmine tõenäosus. Kui panna tulemus graafikule (keskmine tõenäosus vs positiivsete osakaal), siis loodamegi saada võimalikult lähedase tulemuse eelnimetatud ideaalsele [29].

3.1.2.4. Tasakaalustamata andmete probleem

Tasakaalustamata andmed tähendavad, et klasside osakaal andmetes on erinev. Makrosoomia ei ole juba definitsiooni poolest sage nähtus, seega võime ennustavalt öelda, et selle uurimisel tuleb tegelda tasakaalustamata andmetega. Meditsiinis on tasakaalustamata andmed üldse üsna tavaline probleem, sest reeglina ei ole haigeid ja terveid (õnneks) samapalju, vaid terveid on rohkem.

Muidugi saame seda uuringu planeerimisega mõjutada, kuid mitte alati. Sageli kogunevad andmed ravitegevuse käigus, ei valita eelnevalt võrdseid uuritavate ja kontrollrühmi.

Kui andmed on ikkagi märkimisväärselt tasakaalust väljas, st suhe on mitte näiteks 4:6, vaid 1:10, 1:100 jne, siis põhjustab see klasside erinevat jaotust ja näiteks negatiivsete ja positiivsete prognoosimise tõenäosused kujunevad väga erinevaks [30]. Lisaks sellele võimenduvad muud andmete probleemid, näiteks määrangute vead. Kõik see põhjustab proleeme klassifitseerimisel.

Kõige lihtsam võimalus selle probleemiga tegelda on suuremat (sageli negatiivset) klassi juhuslikult vähendada, nii et klassid oleks ligikaudu võrdsed. Sageli kaotame aga niimoodi suure hulga andmeid ja see võib tulemust omakorda mõjutada. Tugevasti tasakaalustamata andmete puhul on sageli paremini toimiv lahendus ühtaegu vähendada suuremat klassi ja paljundada väiksemat, ikka juhuslikkuse alusel [30]. Sellist meetodit kasutatakse ka käesolevas töös.

Tasakaalustamata andmete puhul on suureks abiks ka ristvalidatsioon, sest treeningandmete osadeks jagamisel muutuvad osades ka klasside vahekorrad ja tulemuste keskmistamisel saame parema tulemuse.

3.2. Rasedus, gestatsioondiabeet ja makrosoomia

3.2.1. Makrosoomia, makrosoomne laps

Makrosoomseks võib üldiselt nimetada loodet või vastsündinut, kes on liiga suur, enamasti interpreteeritakse liigset kaalu gestatsioonilise vanuse kohta [31] ja inglise keeles kasutatakse terminit „large for gestational age“ (LGA). Samas, mõned autorid eristavad need mõisted nii, et makrosoomia on suurus olenemata gestatsioonilisest vanusest ja LGA on seotud gestatsioonilise vanusega [32]. Käesolevas töös kasutatakse makrosoomiat pigem just seotuna gestatsioonilise vanusega.

Piiri tõmbamine, kust algab makrosoomia, on eri riikides ja eri allikates erinev. Sageli kasutatav piirväärtus on vastsündinu kaal üle 4000 g [33],[34], kuid kasutatakse ka lävendeid 4500 g ja 5000

g [31]. Samas, Eestis peetakse nö tervislikuks vastsündinute kaaluvahemikuks 3500-4500g ehk kuni 4,5 kg raskuseid vastsündinuid ei peeta liiga suurteks [35].

Mõnel pool on kasutusel nn pondoraalne indeks, mis nagu kehamassiindekski iseloomustab kaalu ja pikkuse vahekorda ja seega iseloomustab paremini kehaehitust ja võimalikku rasvumist kui üksnes kaal. Üks kasutatavatest valemitest on [36]:

$$\frac{vastsündinu\ kaal\ g * 100}{(vastsündinu\ pikkus\ cm)^3}$$

Igapäevapraktikas selle kasutamine siiski levinud ei ole.

Tänapäeval arvatavasti kõige levinum meetod defineerida makrosoomiat on kasvukõverate järgi leida need vastsündinud, kelle kaal on vastavalt gestatsioonilisele vanusele üle 0,9 kvantiili [37], [10]. Selleks on arvutatud rahvusvahelised standardid [10], [38], aga rakendatakse ka Eesti laste põhjal arvutatud kasvukõveraid [39]. Käesolevas töös kasutatakse Euroopa laste kasvukõveraid [10], kuid kuna põhjamaades on vastsündinud üldiselt suuremad, siis kasutame 0,9 kvantiili asemel 0,97 kvantiili.

Makrosoomia toob kaasa probleeme eelkõige seoses sünnitusega, põhjustades probleeme nii emale kui lapsele. Makrosoomia tagajärgedeks võivad olla keisrilõike või vaakumsünnituse vajadus normaalse vaginaalse sünnituse asemel, sünnituse liigne pikenemine, sünnitusteede ja sünnitrauma: lahkliha ulatuslikud rebendid, õlgade düstookia, asfüksia, vastsündinu respiratoorne distress, hüpoglükeemia, ja muud tüsistused [10], [31],[33].

Makrosoomia võimalike põhjustena on esile toodud eelkõige suurt kaaluiivet raseduse ajal , GDM-i ja ema raseduseelset ülekaalu [37],[40], [41]

3.2.2. Gestatsioondiabeet

Gestatsioondiabeet (GDM, *gestational diabetes mellitus*) on süsivesikute ainevahetushäire, mis tekib raseduse ajal (tavaliselt raseduse teises pooles) ja taandub pärast sünnitust. Selle puhul on enamasti veresuhkru tase ainult veidi normist kõrgem, olles madalam kui tavalise diabeedi puhul. [4]

GDM riskirühma, keda peaks testima kasutades glükoositolerantsuse testi (GTT) kuuluvad ülekaalulised naised (KMI >25kg/m^2); rasedad, kellel on esinenud GDM eelmise raseduse ajal; rasedad, kelle esimese astme sugulastel (ema, isa, õde, vend) esineb diabeeti; rasedad, kes on varem sünnitanud suurekaalulise lapse (>4500g) ja rasedad, kellel esineb polütsüstiliste munasarjade sündroom. GDM-ist on enam ohustatud ka rasedad, kelle paastusuhkru tase on 5,1 mmol/L ja enam, kellel esineb polühüdramnion, liigne kaaluiive, glükosuuria käesoleva raseduse ajal [4],[42],[6].

GDM võib põhjustada loote makrosoomiat ning seoses sellega probleeme sünnitusel (meditsiinilise sekkumise vajadust), õlgade düstookiat ja sünnitrauma riski. Emadel esineb sagedamini preeklampsiat ning vastsündinutel hüpoglükeemiat. [4], [40], [42]

3.3. Gestatsioondiabeedi ja makrosoomia varasemad uuringud masinõppe meetoditega

Terviseinformaatika tegeleb tõenäosusliku informatsiooni kasutamisega otsuste tegemiseks [43]. See on edukam koos masinõppega.

Gestatsioondiabeedi ja masinõppe ühendamine, GDM prognoosimine ei ole varasemalt väga suurt tähelepanu leidnud, suur hulk töid on seotud (tavalise) II tüüpi diabeedi ennustamisega ja sellega on tegeldud põhjalikumalt [32]. Samas viimastel aastatel on ilmunud mitmeid töid [44],[45],[46],[47],[48].

Qiu jt (2017) töö [48] on esimene, mis põhineb elektroonilisel tervise andmebaasil (Hiinas) ja katsetab sellel masinõppe meetodeid.

Ka Artzi jt (2020) töö [45] põhineb elektroonilistest terviseregistrist saadud väga suurel andmestikul (üle poole miljoni raseduse, aastatest 2010-2017). Algselt kasutati seitset binaarset tunnust USA-s välja töötatud küsimustiku alusel [49], enamikku neist kasutakse ka käesolevas töös, näiteks ülekaal, vanus, diabeet sugulastel, probleemid varasemate raseduste ajal jms. Paraku sellega nad väga head tulemust ei saavutanud (täpsus 30%), kuid kui nad täiendasid seda elektroonilisest terviseregistrist võetud koguni 2355 tunnusega, tulemused paranesid oluliselt [45].

Nende tulemuste alusel töötati välja uus, parandatud küsimustik, millega saadi ligilähedane tulemus suure hulga tunnuste kasutamisele. Kõige parema tulemuse andiski neil üldtuntud riskitegurite ja laboratoorsete analüüside valitud parameetrite kombinatsioon, kusjuures nad leidsid, et funktsionaalne on ka eelmise raseduse GTT tulemus. Selline tulem viitab võimalustele ehk ka Eestis rohkem kasutada digilugudesse akumuleeritud andmeid, loomulikult tagades isikuandmete kaitse.

Ye jt (2020) ning Wu jt (2020) tööd [44],[47] teostati samuti Hiinas ja mõlemas uuriti erinevate masinõppe meetodite efektiivsust GDM ennustamisel. Kõige paremini toimisid vastavalt gradient-võimendatud otsutuspuu (*Gradient Boosted Decision Tree*, GBDT) ja sügavad närvivõrgud (*Deep Neural Networks*, DNN). Katsetati ka erinevaid tunnuste komplekte, taas andis suurem hulk tunnuseid parema tulemuse [47]. Ye jt töös olid olulisemad tunnused ema KMI, paastusuhkru tase, glükohemoglobiini tase ja triglütseriidide tase.

Ka makrosoomia ennustamisega masinõppe meetoditega pole eriti palju tegeldud ning siingi on kiire areng seotud viimaste aastatega [14],[32],[50],[51],[52],[53]. Tööd põhinevad USA, Hiina ja Jaapani andmetel. Kohati on tööd lausa vastandliku suunitlusega, Ye jt (2020) püüavad parandada makrosoomia ennustamist ultraheli-piltide abiga, samas kui Shigemi jt (2018) püüavad seda teha kasutades emaga seonduvaid tunnuseid ilma ultraheli-piltide abita [32],[51]. Akhtari jt töörühm on kasutanud nii erinevaid emaga seotud tunnuseid [50],[52] kui ka keskendunud biokeemilistele markeritele [53]. Kasutatud on erinevaid masinõppe meetodeid (LR, NB, RF, SVM jt) ja ka spetsiifiliselt välja töötatud meetodeid. Shigemi jt töös [51], kus kasutati juhuslikku metsa, on täpsus üllatavalt madal (0,03-0,04) ja ka tundlikkus keskmine (0,6-0,88), siiski AUC on 0,88, mis on päris hea, veel rõhuvad nad kõrgele negatiivsete ennustusvõimele (0,96-1,0), kuid makrosoomia puhul pole see praktikas nii oluline, pigem oleks tähtis leida üles makrosoomsed. Teistes töödes, eriti Ye jt artiklis [32] on kasutatud rohkem erinevaid mudeleid ja AUC/õigsus kõigub 0,7-0,98 vahel.

GDM-i ja makrosoomia uuringud on viimastel aastatel kiires arengus ja ilmunud on erinevaid uuringuid, kuid tegu on peamiselt regionaalsete uuringutega ja Euroopas on töid ilmunud vähe. Meie töö eesmärgiks on katsetada GDM-i ja makrosoomia ennustamist Eesti andmetel, vaadelda tulemuste erinevust/sarnasust teiste uuringutega, vaadelda riskitegurite/tunnuste väärtust prognoosimisel. Uuenduslik aspekt on GTT rolli analüüsimine makrosoomia ärahoidmisel.

3.4. Probleemi sõnastus, eesmärgid

Arstide eesmärk on võimaluste piires vältida loote kasvu liiga suureks ehk makrosoomia teket, kuna sellega kaasnevad mitmed ohud sünnitusel nii emale kui lapsele. Seetõttu oleks oluline riskitegurite alusel ennustada võimalikult raseduse algul makrosoomia tekke võimalust ja võtta kasutusele meetmed selle ärahoidmiseks (GDM testimine ja ravi, dieet jm). Käesolevas töös on eesmärgiks ennustada GDM ja makrosoomia teket, analüüsida makrosoomia ja GDMi seost ning hinnata riskitegurite olulisust kasutades erinevaid andmeteaduse ja masinõppe meetodeid.

Täpsemad eesmärgid on:

1. Selgitada, milliste riskiteguritega on seotud gestatsioondiabeet.
2. Selgitada, milliste riskiteguritega on seotud makrosoomia.
3. Leida makrosoomia seosed sünnitusaegsete probleemidega nagu õlgade düstookia, perineumi rebendid.
4. Prognoosida GDM-i teket riskitegurite alusel.
5. Prognoosida makrosoomiat raseduse esimeses pooles määratavate riskitegurite alusel.
6. Prognoosida makrosoomiat hilisemate tunnuste: GDM, rasedusaegse kaaluiibe ja teiste seonduvate riskitegurite alusel.
7. Analüüsida GDM, GTT ja teiste tunnuste rolli makrosoomia prognoosimisel.
8. Visualiseerida saadud tulemusi.

4. Materjal ja metoodika

4.1. Andmed

Uuringuandmed on kogutud kolmel ajaperioodil:

Andmestik 1: Uuringu „Gestatsioondiabeedi sõeluuring Tartu Ülikooli Kliinikumis 2012 aastal“ raames koguti andmed GDMI riskitegurite esinemise kohta 2012. a SA TÜ Kliinikumi naistenõuandla perekeskuses rasedusega arvel olnud naistelt GDM kontroll-lehe abil. Rasedat jälgiv ämmaemand täitis kontroll-lehe raseduse jooksul, andmed sisestati andmetabelisse ning haiguslugudest leiti rasedustulem. Uuringu tulemused on avaldatud artiklis [4]. Andmestik sisaldab 1073 raseduse kirjet.

Andmestik 2: Uuringu „Inimese viljakuse ja raseduse kuluga seotud mitte-invasiivsete biomarkerite arendamine“ (akronüüm Happy Pregnancy, HP) raames kaasati SA TÜ Kliinikumi naistenõuandla rasedusaegsele jälgimisele tulnud naised ajavahemikus 01.09.2012-31.08.2015. Andmed eelnevate raseduste, üldiste terviseandmete, eluviiside, rasedusaegsete vaevuste kohta koguti küsimustike ja elektroonilise rasedakaardi abil. Andmed raseduse lõppe kohta saadi meditsiinilisest dokumentatsioonist. Andmestik sisaldab 2248 raseduse kirjet.

Andmestik 3: 2018. a SA TÜK naistekliinikus rasedusaegsel jälgimisel olnud rasedate andmed koguti päringu teel kliinikumi elektroonilisest haiguslugudest. Andmestik sisaldab 1768 raseduse kirjet.

Uurimistööks kõikide nende andmete baasil on taotletud TÜ eetikakomitee luba seoses sellega, et tegu on inimuuringutega, luba ka anti, protokolli number: 291/T-3 18. märtsist 2019. a (lisa 1). Uuritavate isikuandmed olid andmestikust eraldatud ja ei olnud analüüsijale kättesaadavad. Uuritavad olid tähistatud koodide või järjekorranumbritega.

Analüüsiks kasutati Happy Pregnancy (HP) andmestikku eraldi ja kolme andmestikku kokkuliidetuna (ühiseid tunnuseid). Samade tunnuste puhul oli nende hindamine patsientidel ühesugune ja seega võime kasutada andmestikke koos.

4.2. Kasutatud tunnused

Tunnused valiti välja lähtuvalt andmestikust 2 (Happy Pregnancy), kus leiduvast suurest hulgast tunnustest (>90) valiti välja käesolevaks uuringuks vajalikud tunnused, mis on arstide hinnangul seotud makrosoomia ja rasedusdiabeediga. Hiljem, andmestike liitmisel jäid osad tunnused neist siiski koondandmestikust välja, sest puudusid mõnes andmestikus. Mõningaid tunnuseid siiski analüüsiti ka ainult andmestiku 2 põhjal. Valitud ja kasutatud tunnused on näidatud tabelis 1.

Tabel 1. Kasutatud tunnused

Tunnus	Tüüp	Ühik/ seisundid	Andmestik	Riskiteguri hindamise aeg raseduse jooksul	Tulem T/ faktor F
GDM ja makrosoomia riskitegurid (teadaolevad ja arvatavad)					
Diabeet I ringi sugulastel	binaarne	0,1	koond	alguses	F
Varasema raseduse ajal GDM	binaarne	0,1	HP	alguses	F
Varasema raseduse ajal makrosoomne laps	binaarne	0,1	koond	alguses	F
Paastusuhkur (veresuhkur tühja kõhuga)	binaarne	<5.1 mmol/l (0), >= 5.1 mmol/l (1)	koond	alguses	F
PCOS	binaarne	0,1	koond	alguses	F
Mitmes sünnitus	diskreetne arvtunnus	1-15	HP	alguses	F
Ema kaal	pidev arvtunnus	kg	koond	alguses	F

Tunnus	Tüüp	Ühik/ seisundid	Andmestik	Riskiteguri hindamise aeg raseduse jooksul	Tulem T/ faktor F
Ema vanus	pidev arvtunnus	aasta	koond	alguses	F
KMI	pidev arvtunnus/ arvutatud	kg/m^2	koond	alguses	F
Polühüdramnion	binaarne	0,1	koond	keskel	F
Lapse sugu	nominaalne	poiss, tüdruk (1,2)	koond	keskel/lõpus	F
Rasedusaegne kaaluiive	pidev arvtunnus	kg	HP	lõpus	F
GDMi hinnang riskitegurite alusel					
madal	nominaalne	ei tehtud GTT (1)	koond	keskel	F
kõrge		ei tehtud GTT (2), GTT tehtud, kuid negatiivne GDM suhtes (3), GTT tehtud ja diagnoositud GDM (4)			
Raseduse tulem					
GDM	binaarne	0,1	koond	keskel	F/T
Preeklampsia	binaarne	0,1	HP	lõpus	T
Raseduse kestus sünnituse ajal	diskreetne arvtunnus	päeva	koond	lõpus	F
Sünnitus	nominaalne	normaalne vaginaalne,	koond	lõpus	T

Tunnus	Tüüp	Ühik/ seisundid	Andmestik	Riskiteguri hindamise aeg raseduse jooksul	Tulem T/ faktor F
		vaakum, keisrilõige (1,2,3)			
Sünnikaal	pidev arvtunnus	g	koond	lõpus	T
Sünnikaalu Z-väärtus	pidev arvtunnus/ arvutatud		koond	lõpus	T
Makrosoomne laps	binaarne/arvutatud	0,1	koond	lõpus	T
Õlgade düstookia	binaarne	0,1	HP	lõpus	T
Perineumi rebend (III-IV aste)	binaarne	0,1	HP	lõpus	T
Lisatunnused					
Uuritava kood	nominaalne		koond		
Valim	nominaalne	2012 (1), HP (2), 2018 (3)	koond		

4.3. Andmeanalüüs

4.3.1. Python ja selle teegid

Andmeanalüüs teostati peamiselt programmeerimiskeelt Python kasutades (versioon 3.6.9) [54] keskkonnas Google Colaboratory. Andmete ettevalmistamisel kasutati ka programmipaketti MS Excel (Office 365).

Pythoni teekidest kasutati järgmisi: NumPy [55], Pandas [56], SciPy [57], Scikit-Learn [58], Matplotlib [59] ja Seaborn [60]. Nimetatutest on NumPy ja Pandas sobivad erinevate

andmestruktuuride haldamiseks, SciPy ja Scikit-Learn andmeanalüüsi funktsioonideks ning Matplotlib ja Seaborn visualiseerimiseks.

Koodi kirjutamisel kasutati teekide API-sid ja lisaks mõnikord ka muid allikaid. Vastavad viited on ära toodud koodide juures. Koodid on üles laetud ja kättesaadavad GitHubis: https://github.com/silviapihu/magistritoo.

4.3.2. Andmeanalüüs

Uuritavateks tunnusteks kasutati peamiselt GDMi ja makrosoomiat. Prooviti ka erinevate tunnuste korrelatsiooni sünnikaalu ja selle Z-väärtusega, kuid korrelatsioonid olid väga nõrgad või puudusid, seetõttu loobuti ka edasisest regressioonanalüüsist ja keskenduti klassifikatsioonile. Tegelikult on ka meditsiiniliselt olulisem ennustada mitte täpset sünnikaalu, vaid seda, kas laps on liiga suur (makrosoomne) või mitte ja võivad tekkida probleemid sünnitusel.

Sõltumatute tunnuste ja GDM/makrosoomia vaheliste seoste kindlakstegemiseks kasutati Studenti t-teste pidevate arvtunnuste jaoks, $\chi 2$-teste ja Fisheri täpset testi, mis töötavad sagedustabelite alusel, kategooriatunnuste (binaarsete ja nominaalsete) jaoks. Fisheri testi alusel arvutati šansside suhted, mis näitavad kui palju suureneb GDM/makrosoomia tõenäosus vastavale tunnusele eksponeeritutel. Arvutati tunnuste kaupa välja rühmade (GDM on/ei ole; makrosoomia on/ei ole) keskmised pidevate arvtunnuste jaoks ning positiivsete osakaal rühmades binaarsete tunnuste jaoks.

Kuna andmed olid mittetasakaalulised, siis kasutati enne klassifitseerimist suurema (negatiivse) klassi juhuslikku vähendamist ja väiksema (positiivse klassi) juhuslikku paljundamist.

Klassifitseerimiseks GDM/makrosoomia alusel kasutati järgmisi mudeleid: lähimate naabrite meetod (KNN), otsustuspuu (DT), Gaussi naiivne Bayesi meetod (GNB), logistiline regressioon (LR), juhuslik mets (RF), tugivektormasin (SVM) ja lineaarne diskriminantanalüüs (LDA). Logistilist regressiooni kasutati nii lineaarsena kui ka polünomiaalsete tunnustega (teises astmes).

Klassifikatsioonimudelite hindamiseks kasutati nn segadusmaatriksit ja sellest arvutatud järgmisi näitajaid: õigsus, täpsus, tundlikkus, spetsiifilisus, valepositiivsete määr, valenegatiivsete määr, F1 ja ROC-kõvera alune pindala.

Parimate mudelite otsinguks kasutati mudelite võrdlemisel ristvalideerimise meetodit (10 lahknemisega) ja selle põhjal arvutatud keskmist täpsust ning mudelite kalibreerimist.

Mudeli koefitsientide põhjal hinnati tunnuste osakaalu ennustamisel. Kasutati ka tunnuste rekursiivset eemaldamist, et kindlaks teha optimaalne tunnuste komplekt.

5. Tulemused

5.1. Riskitegurite ja rasedustulemite seos gestatsioondiabeediga

Rasedad, kellel oli GTT alusel diagnoositud GDM ja nendel, kellel GDM ei olnud diagnoositud tunnuste keskmiste väärtuste võrdlemise (Studenti t-test, arvtunnuste puhul) ja sagedustabelite $\chi 2$-testide põhjal (binaarsed ja nominaalsed tunnused) leitud seosed on toodud tabelis 2. Rasedaid, kellel GTT ei teostatud puuduva näidustuse tõttu (nn madal risk) oli 2303, riskiteguritega rasedatest oli neid, kellel jäeti GTT tegemata vaatamata näidustuse olemasolule, 944. GTT andis normaalse (negatiivse) tulemuse 1340 naisel ja GDM diagnoositi 430 rasedal. Tulemused kinnitavad varasemalt teadaolevat infot, et GDM-iga rasedatel esineb sagedamini haiguse riskitegureid (kõrge KMI, paastusuhkur üle normi, varasema raseduse ajal GDM või suur laps jne) ning mitmeid raseduse ja sünnitusega seonduvaid probleeme: makrosoomia (19,3% versus 10,8%), polühüdramnion (3,9% versus 1,8%), operatiivne sünnitus (31,2% versus 20%). Huvitav asjaolu on, et GDM-diagnoosiga rasedatel on kaaluiive oluliselt väiksem kui teistel.

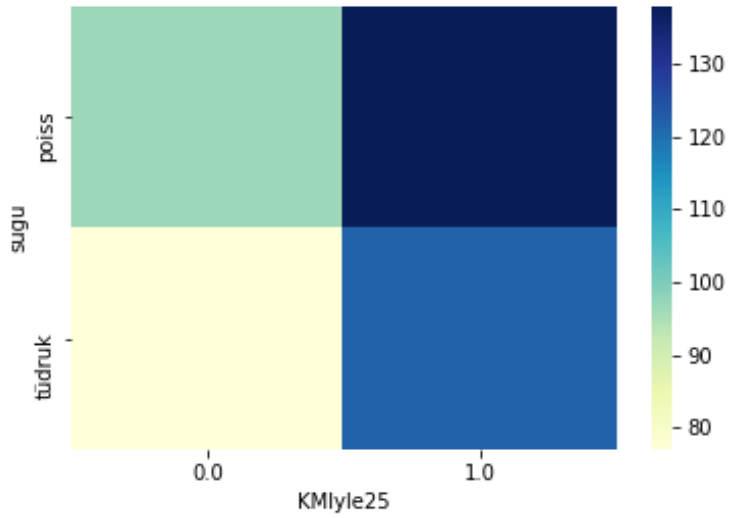
Statistiliselt mitteolulised olid GDM seosed lapse sünnikaaluga, sellega, mitmenda sünnitusega oli tegu, lapse soo, preeklampsia, õlgade düstookia ja perineumi rebendiga.

Värvikaardilt tuleb hästi välja ema raseduseelse kehamassiindeksi seos GDM-iga (joonis 1). Samast ilmneb ka erinevus vastavalt lapse soole.

Tabel 2. GDM-i seosed erinevate riskitegurite ja raseduse lõppega. GDM diagnoositud ja GDM mitte diagnoositud rasedatele on antud tunnuste keskmine väärtus ± standardhälve pidevate tunnuste korral, osakaal binaarsete tunnuste korral. HP: tunnused, mis puudusid koondandmestikus ja mida on analüüsitud ainult Happy Pregnancy andmete põhjal. Studenti t-testi või $\chi 2$ Pearsoni statistiku olulisuse tõenäosus vastavalt Bonferroni korrektsioonile 20 tunnuse kohta: * < 0,0025, **< 0,0005.

	Rasedad, kel diagnoositud GDM n=435	Rasedad, kel ei ole GDM diagnoositud n=4352	Efekt	Statistik	*p*-väärtus	And-mestik
Pidevad tunnused			Keskmiste vahe %	t-väärtus		
Ema vanus (aastates)	31,0±5,1	29,2±5,3	6,0%	6,79	**	koond
Ema kaal (kg)	76,9±17,0	65,1±12,5	18,1%	18,37	**	koond
Ema KMI (kg /m^2)	27,4±6,0	23,2±4,2	18,1%	19,60	**	koond
Raseduse kestus sünnituse ajal (päeva)	275±12	278±12	1,0%	-5,17	**	koond
Sünnikaal	3614±548	3549±534	1,8%	40,18	**	koond
Sünnikaalu Z-väärtus	0,94±0,98	0,66±0,94	30,0%	49,75	**	koond
Rasedusaegne kaaluiive (kg)	11,1±6,2	13,0±4,8	17,1%	4,01	**	HP
Kategooria-tunnused			Erinevus kordades	χ2-statistik		
Ema KMI > 25kg/m^2	59,9%	24,7%	2,4	242,51	**	koond
Diabeet sugulastel	19,1%	8,4%	2,3	52,94	**	koond
Varasema raseduse ajal laps >4500g	7,8%	1,6%	4,9	70,76	**	koond
PCOS	3,7%	0,9%	4,1	24,15	**	koond
Paastusuhkur üle normi	43,0%	12,8%	3,4	273,20	**	koond
Makrosoomia	19,3%	10,8%	1,8	27,15	**	koond

Sünnitusviis operatiivne	31,2%	20,0%	1,6	34,79	**	koond
Polühüdramnion	3,9%	1,8%	2,2	8,15	*	koond
Varasema raseduse ajal GDM	12,4%	0,7%	17,7	99,38	**	HP



Joonis 1. GDM-i diagnoosiga rasedate arvu jagunemine vastavalt sündinud laste soole (y-teljel) ja ema raseduseelsele KMI-le (üle 25 või mitte, x-teljel). Skaala paremal näitab GDM-iga emadel sündinud laste arvu, kes vastavasse kategooriasse kuulusid ja sellele vastavat värvitooni,.

5.2. Riskitegurite ja rasedustulemite seos makrosoomiaga

Makrosoomse (n=580, 11,5%) ja mittemakrosoomse (n=4437; 88,5%) lapse sünniga lõppenud raseduste ja emade tunnuste keskmiste väärtuste võrdlemise (Studenti t-test, arvtunnuste puhul) ja sagedustabelite χ2-testide põhjal (binaarsed ja nominaalsed tunnused) leitud seosed on toodud

tabelis 3. Mittemakrosoomsete laste hulka arvati nii normaalkaalus vastsündinud kui ka need, kellel sünnikaal oli alla 0,1 kvantiili (rasedusekestuse kohta väike laps). Sarnaselt GDM-iga on ka makrosoomse lapse sünniga lõppenud raseduste korral emadel kõrgem KMI, paastusuhkur üle normi jne. Makrosoomsetel lastel esineb rohkem GDMi kui mittemakrosoomsetel.

Erinevalt GDMist esines makrosoomiaga lapse sünnitustel ajal oluliselt enam õlgade düstookiat (2,6% vs 0,5%) ja perineumi III ja IV astme rebendeid (2,2% vs 1,1%).

Statistiliselt mitteolulisteks osutusid makrosoomia seosed ema vanuse, PCOSi, varasema raseduse ajal GDM-i esinemise, sugulastel esineva diabeedi, raseduse kestuse, preeklampsia ja õlgade düstookiaga.

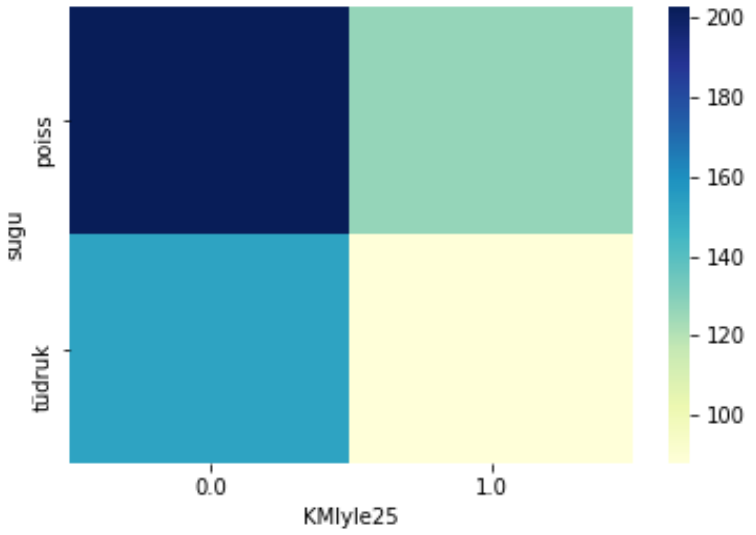
Üksikute tunnuste analüüsimisel värvikaartide abiga tulevad välja erinevused makrosoomsete poisslaste ja tütarlaste emade tunnustes, näiteks kehamassiindeksi puhul (joonis 2). Huvitav tulemus tuleb GDM riskigruppide võrdlemisel, kus kõige enam makrosoomseid lapsi sündis rasedate rühmas, kellel oli mingi riskifaktor ja keda testiti GDM suhtes, kuid kes said vastuse, et neil ei ole gestatsioondiabeeti (joonis 3). Tulid välja ka erinevused poiste ja tüdrukute vahel.

Viiuldiagrammilt tuleb välja sünnikaalu seos varasema üle 4,5 kg kaalunud lapse sünniga (joonis 4) – varem suure lapse sünnitanud naistel sünnib sagedamini uuesti suurema sünnikaaluga laps. Nähtub ka erinevus erinevast soost laste vahel, just poisslaste puhul avaldub efekt tugevamini.

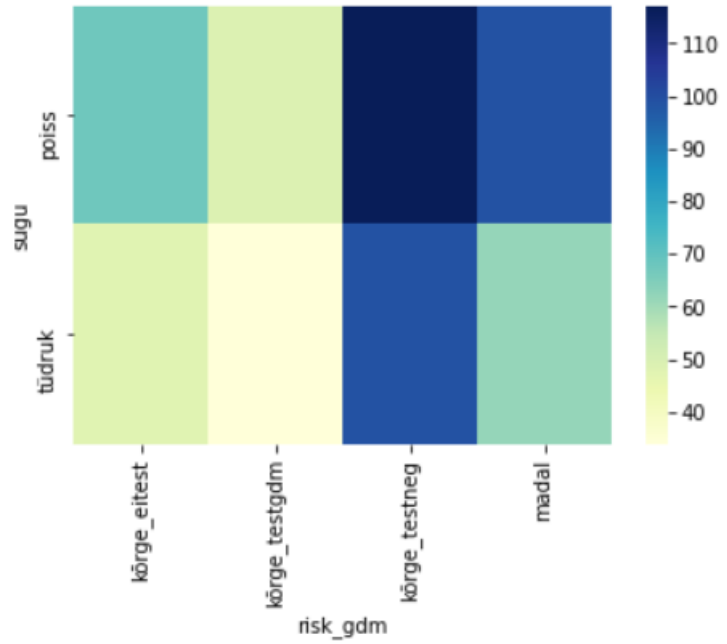
Tabel 3. Makrosoomia seosed erinevate riskitegurite ja raseduse lõppega. Makrosoomsetele ja mittemakrosoomsetele lastele on antud tunnuste keskmine väärtus ± standardhälve pidevate tunnuste korral, osakaal binaarsete tunnuste korral. HP: tunnused, mis puudusid koondandmestikus ja mida on analüüsitud ainult Happy Pregnancy andmete põhjal. Studenti t-testi või $\chi 2$ Pearsoni statistiku olulisuse tõenäosus vastavalt Bonferroni korrektsioonile 20 tunnuse kohta: * < 0,0025, **< 0,0005

	Makro-soomsed vastsündinud n=580	Mitte-makrosoomsed vastsündinud n=4437	Efekt	t-testi või $\chi 2$ statistik	*p*-väärtus	And-mestik
Pidevad tunnused			Keskmiste vahe %	t-väärtus		
Ema kaal (kg)	71,5±15,3	65,4±12,9	9,3 %	10,347	**	koond
Ema KMI (kg /m^2)	24,8±5,1	23,4±4,4	5,9 %	7,111	**	koond
Vastsündinu sünnikaal (grammi)	4294±296	3458±482	24,2%	40,18	**	koond
Rasedusaegne kaaluiive (kg)	14,3±5,8	12,7±4,8	12,5 %	3,948	**	HP
Mitmes sünnitus	2,1±0,9	1,8±0,9	16,7 %	4,144	**	HP
Kategooria-tunnused			Erinevus kordades	$\chi 2$-statistik		
Ema KMI > 25kg/m^2	37,4%	26,5%	1,4	29,29	**	koond
Varasema raseduse ajal laps >4500g	7,6%	1,4%	5,4	93,12	**	koond
Paastusuhkur üle normi	19,5%	14,9%	1,3	8,64	*	koond
Polühüdramnion	5,0%	1,6%	3,1	30,48	**	koond

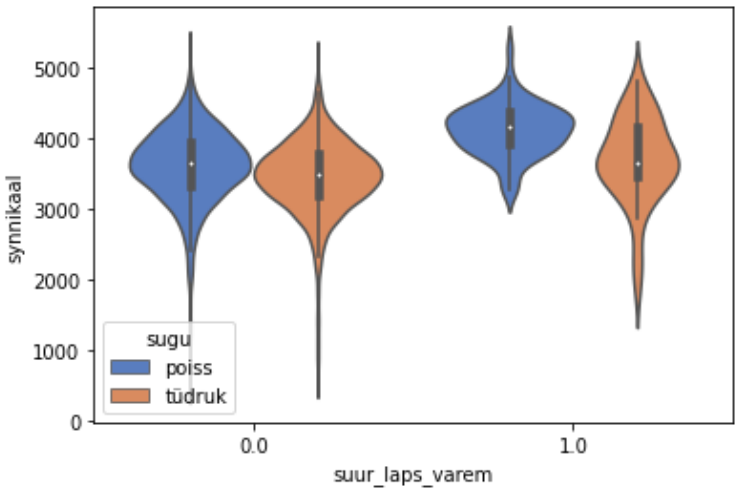
	Makro-soomsed vastsündinud n=580	Mitte-makrosoomsed vastsündinud n=4437	Efekt	t-testi või $\chi 2$ statistik	*p*-väärtus	And-mestik
Kuuluvus GDM riskigruppi	- (gruppe >2)	-		102,19	**	koond
GDM	14,5%	7,9%	1,8	27,15	**	koond
Sünnitusviis operatiivne	29,6%	20,0%	1,5	40,56	**	koond
Vastsündinu sugu (poiss)	57,8%	51,0%	1,1	8,12	*	koond
Perineumi rebend (III-IV aste)	2,2%	1,2%	1,8	40,56	**	HP



Joonis 2. Makrosoomsete laste arv vastavalt sündinud laste soole (y-teljel) ja ema kehamassindeksile enne rasedust (üle 25 või mitte, x-teljel). Skaala paremal näitab makrosoomsete laste arvu, kes vastavasse kategooriasse kuulusid ja sellele vastavat värvitooni,.



Joonis 3. Sündinud makrosoomsete laste arv vastavalt laste soole (y-teljel) ja GDM riskirühmale (x-teljel). Kategooriad x-teljel: madal - madal risk, puudus vajadus testida, kõrge- riskitegurid olemas, vajalik testida, nende hulgas eitest - test jäeti tegemata, testgdm – testiti ja diagnoositi GDM, testneg -testiti ja ei leitud GDMi. Skaala paremal näitab makrosoomsete laste arvu, kes vastavasse kategooriasse kuulusid ja sellele vastavat värvitooni,.



Joonis 4. Sünnikaalu (y-teljel) jaotused varem üle 4,5-kilogrammise lapse sünnitanud naistel ja ülejäänutel (x-teljel) lapse soo kaupa (värvikood legendil).

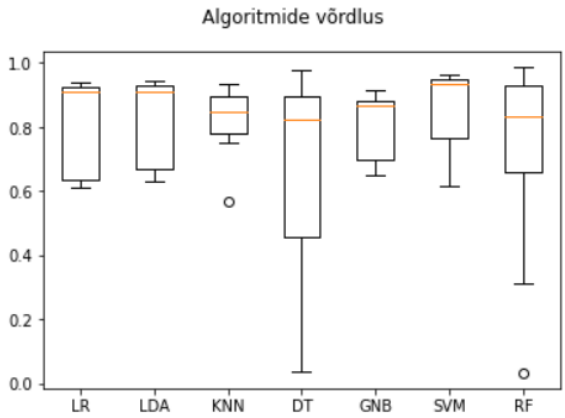
5.3. Gestatsioondiabeedi prognoosimine riskitegurite alusel

Tunnused - riskitegurid, mida kasutati GDM prognoosimisel ja šansside suhted nende jaoks on toodud tabelis 4. Šansside suhe näitab, kui mitu korda erineb uuritava sündmuse toimumise šanss faktoritele eksponeerituil võrreldes mitteeksponeeritutega. Kõige suurem šanss GDM-i haigestumiseks on naistel, kellel on varem sündinud suur laps ja kellel on paastusuhkru näit normist kõrgem, aga väga olulisteks soodustavateks teguriteks on ka ema kõrge KMI enne rasedust ja polütsüstiliste munasarjade sündroom.

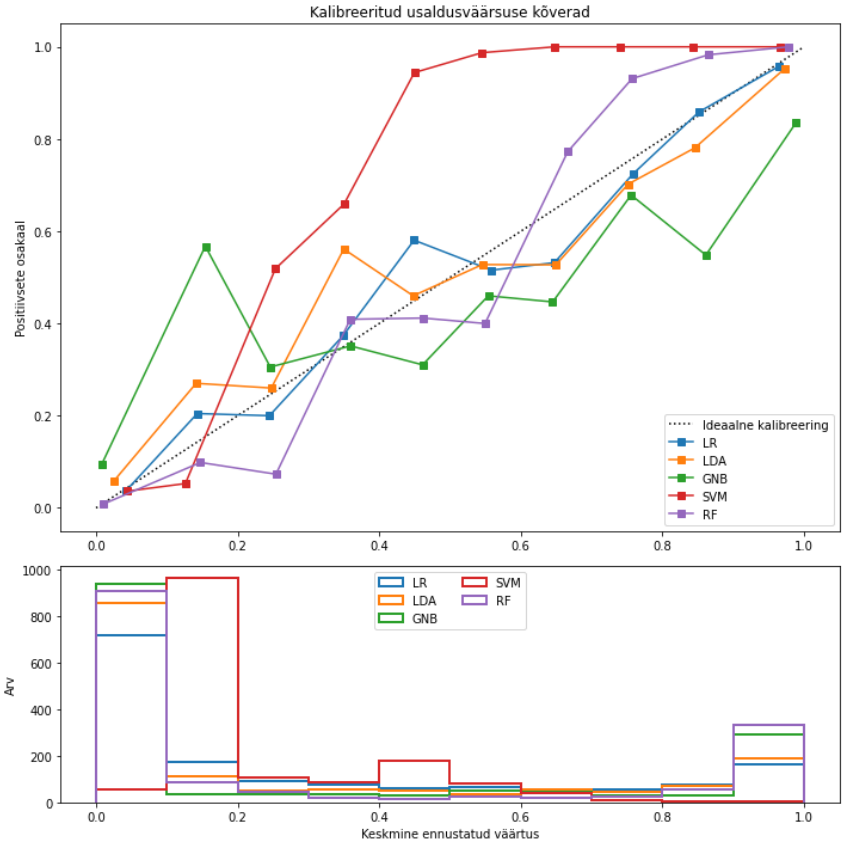
Tabel 4. Šansside suhted riskifaktoritele eksponeeritutel GDM suhtes Fisheri täpse testi alusel.

Tunnus	Šansside suhe	p-väärtus
Varasema raseduse ajal laps >4500g	5,24	<0,0001
Paastusuhkur üle normi	5,15	<0,0001
Ema KMI > 25kg/m^2	4,55	<0,0001
PCOS	4,13	<0,0001
Diabeet sugulastel	2,59	<0,0001
Polühüdramnion	2,23	0,0057
Vanus üle 40	1,30	0,3108

Erinevate binaarse klassifikatsiooni mudelite võrdlused GDM-i prognoosimisel on toodud joonistel 5 ja 6.



Joonis 5. Algoritmide õigsuse (y-teljel, kastidel mediaan ja kvartiilid) võrdlus GDM prognoosimisel, saadud ristvalideerimise meetodil. Algoritmide lühendeid vt sõnastikust. Algoritmide puhul on kasutatud vaikimisi parameetreid.



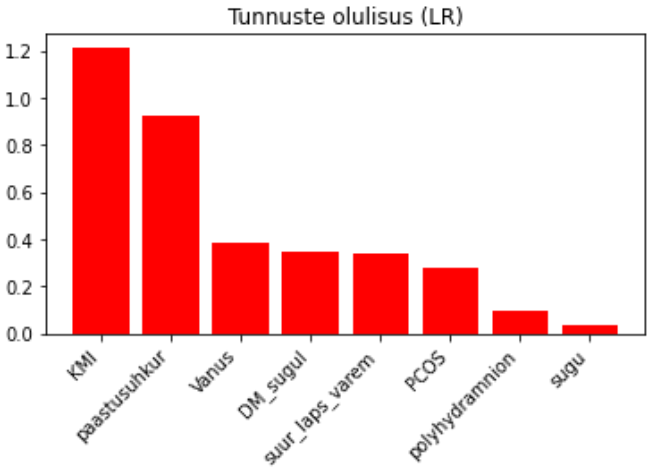
Joonis 6. Vaadeldud ja prognoositud tõenäosuste alusel koostatud kalibreerimisgraafik erinevatele klassifitseerimisalgoritmidele (vaikimisi parameetritega) GDM prognoosimisel. Mudelite lühendid sõnastikus.

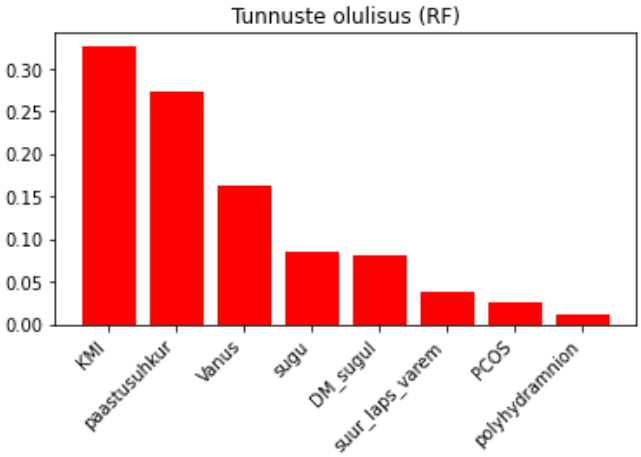
Eeltoodud jooniste alusel valiti edasiseks tööks juhuslik mets, logistiline regressioon ja lineaarne diskriminantanalüüs. KNNi ei valitud, kuna vaatamata suhteliselt headele tulemustele tuleb silmas pidada seda, et kasutasime eeltöötlust, mis paljundas ka positiivset klassi ehk siis lähimate naabrite hulgas võib kergesti leiduda sama juhtumi „kloone“. Võrkotsingu tulemusena leitud parimate mudelite parameetrid on näidatud lisas 2, lineaarse diskriminantanalüüsi jaoks kasutati vaikimisi parameetreid. Nende mudelite hinnangukriteeriumid on toodud tabelis 5. Sellest nähtub, et GDM on kasutatud tunnuste alusel üsna hästi ennustatav, näitajad on väga head. Parim on juhusliku metsa tulemus, kuid peaaegu sama hea ka logistiline regressioon polünoomtunnustega.

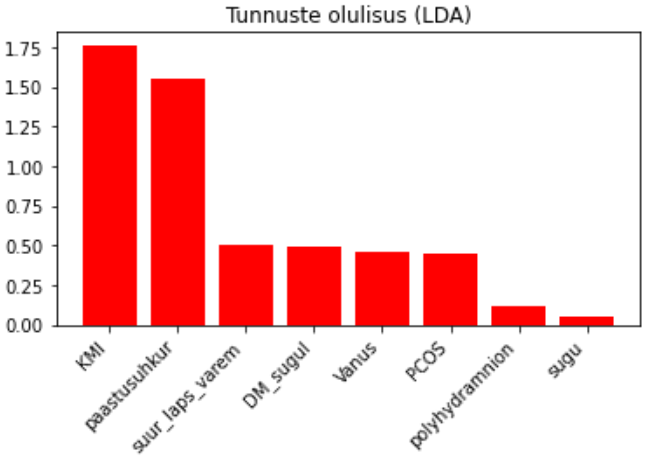
Tabel 5. GDM ennustusvõime hinnangu kriteeriumid (testandmetel, positiivse klassi kohta) logistilise regressiooni (lineaarne ja polünoomtunnustega), lineaarse diskriminantanalüüsi ja juhusliku metsa mudelite jaoks:

Kriteerium/mudel	LR lineaarne	LR polünoomtunnustega	LDA	RF
Valepositiivsete määr (FPR)	0,07	0,03	0,08	0,02
Valenegatiivsete määr (FNR)	0,33	0,17	0,32	0,07
Õigsus (ACC)	0,85	0,93	0,84	0,97
Täpsus (PPV)	0,80	0,93	0,80	0,96
Tundlikkus (TPR)	0,67	0,83	0,68	0,93
Spetsiifilisus (TNR)	0,93	0,97	0,92	0,98
F1	0,73	0,88	0,73	0,94
AUC	0,80	0,90	0,80	0,96

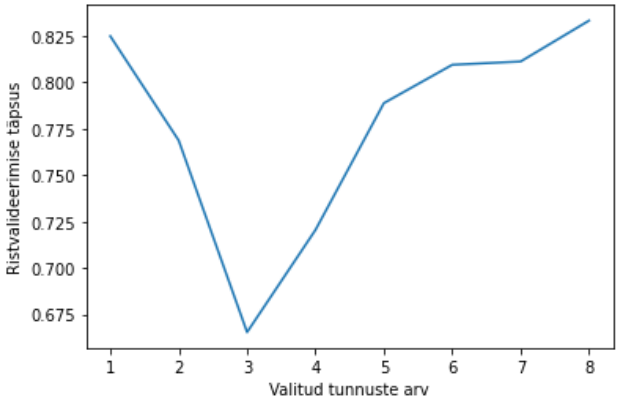
Samade mudelite koefitsientide põhjal arvutatud tunnuste olulisus on näidatud joonisel 7 ja rekursiivselt tunnuste eemaldamise tulemus logistilises regressioonis joonisel 8. Optimaalne tunnuste arv on 8 ja eemaldada ei saa ühtegi tunnust. Kõige olulisemad tunnused kõikide mudelite järgi olid ema KMI enne rasedust ja paastusuhkru näit.







Joonis 7. Tunnuste olulisus GDM-i prognoosimisel (arvutatud mudelite koefitsientide alusel) logistilise regressiooni, juhusliku metsa ja lineaarse diskriminantanalüüsi mudelite puhul.



Joonis 8. Tunnuste rekursiivne eemaldamine ristvalideerimise meetodil juhusliku metsa alusel. Õigsus (y-teljel) vastavalt tunnuste arvule (x-teljel).

5.4. Makrosoomia prognoosimine riskitegurite alusel

Tunnused - riskitegurid, mida kasutati makrosoomia prognoosimisel ja šansside suhted nende jaoks on toodud tabelis 6. Kõige suurem šanss makrosoomse lapse sünniks on rasedatel, kellel on varem sündinud makrosoomne laps, makrosoomiale viitab hästi ka polühüdramnioni teke ning ligi kaks korda suurendab võimalust ka GDM. Ühtegi head esimese raseduse algul esinevat tunnust, mis viitaks makrosoomiale, selle alusel paraku välja ei tulnud.

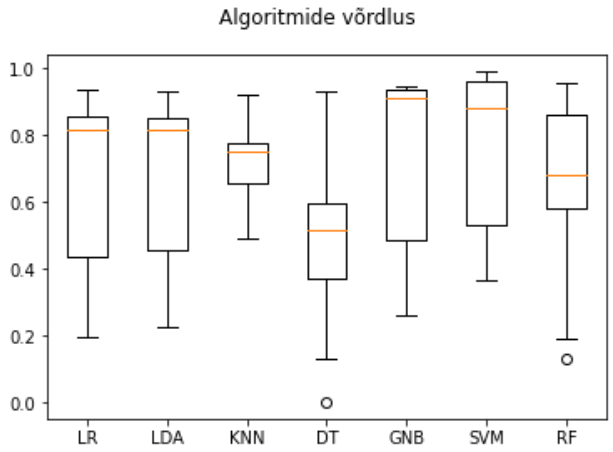
Tabel 6. Šansside suhted riskifaktoritele eksponeeritutel makrosoomia suhtes Fisheri täpse testi alusel.

Tunnus	Šansside suhe	p-väärtus
Varasema raseduse ajal laps >4500g	5,70	<0,001
Polühüdramnion	3,28	<0,001
GDM	1,97	<0,001
Ema KMI > 25kg/m^2	1,65	<0,001
Paastusuhkur üle normi	1,39	0,005

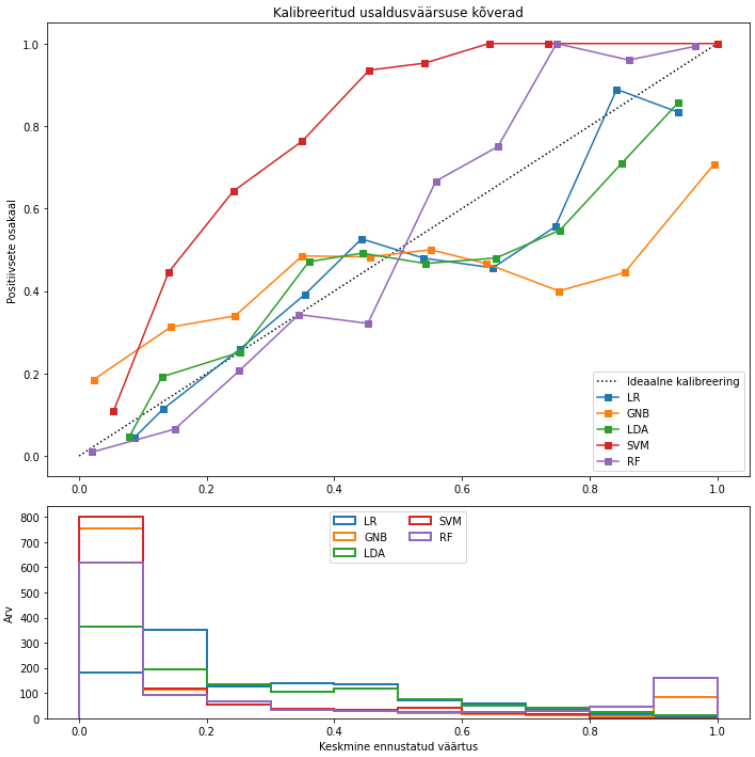
Tunnus	Šansside suhe	p-väärtus
Vanus üle 40	1,13	0,61
Diabeet sugulastel	1,10	0,54
PCOS	1,05	0,84
Lapse sugu (tüdruk)	0,76	0,002

Makrosoomia prognoosimisel klassifikatsioonimudelite alusel kasutati kahte veidi erinevat andmestikku: kõikide koondandmestikus leiduvate riskifaktoritega, mis on teada vähemalt raseduse keskpaigas (samad, mis toodud tabelis 6) ja väiksema tunnuste arvuga andmestik, ainult tunnused, mis on teada juba raseduse alguses: ema KMI, ema vanus, diabeet sugulastel, varasem suure lapse sünd, PCOS ja paastusuhkur. Põhjuseks on asjaolu, et parim oleks makrosoomia võimalust diagnoosida juba võimalikult raseduse alguses, et siis näiteks dieediga püüda vältida makrosoomia tekkimist.

Kõikide võimalike tunnustega tehti taas läbi mudelite valiku protsess (võrreldi mudelite õigsust ristvalideerimise abiga ja kalibreeriti mudelid) ning sõelale jäid needsamad kolm: logistiline regressioon, lineaarne diskriminantanalüüs ja juhuslik mets (joonis 9, joonis 10, tabel 7). Võrkotsingu tulemusena leitud parimate mudelite parameetrid on näidatud lisas 2, lineaarse diskriminantanalüüsi jaoks kasutati vaikimisi parameetreid. Nagu tulemustest nähtub, on makrosoomia prognoosimine raskem kui GDMi prognoosimine ja prognoosimise tulemused logistilise regressiooni ja lineaarse diskriminantanalüüsi abil ei ole nii head, logistilise regressiooni ja lineaarse diskriminantanalüüsi puhul on suur valenegatiivsete määr ja väike tundlikkus. Juhusliku metsa meetodil tehtud prognooside näitajad on siiski väga head.



Joonis 9. Algoritmide õigsuse (y-teljel, kastidel mediaan ja kvartiilid) võrdlus makrosoomia prognoosimisel, saadud ristvalideerimise meetodil. Algoritmide lühendeid vt sõnastikust. Algoritmide puhul on kasutatud vaikimisi parameetreid.

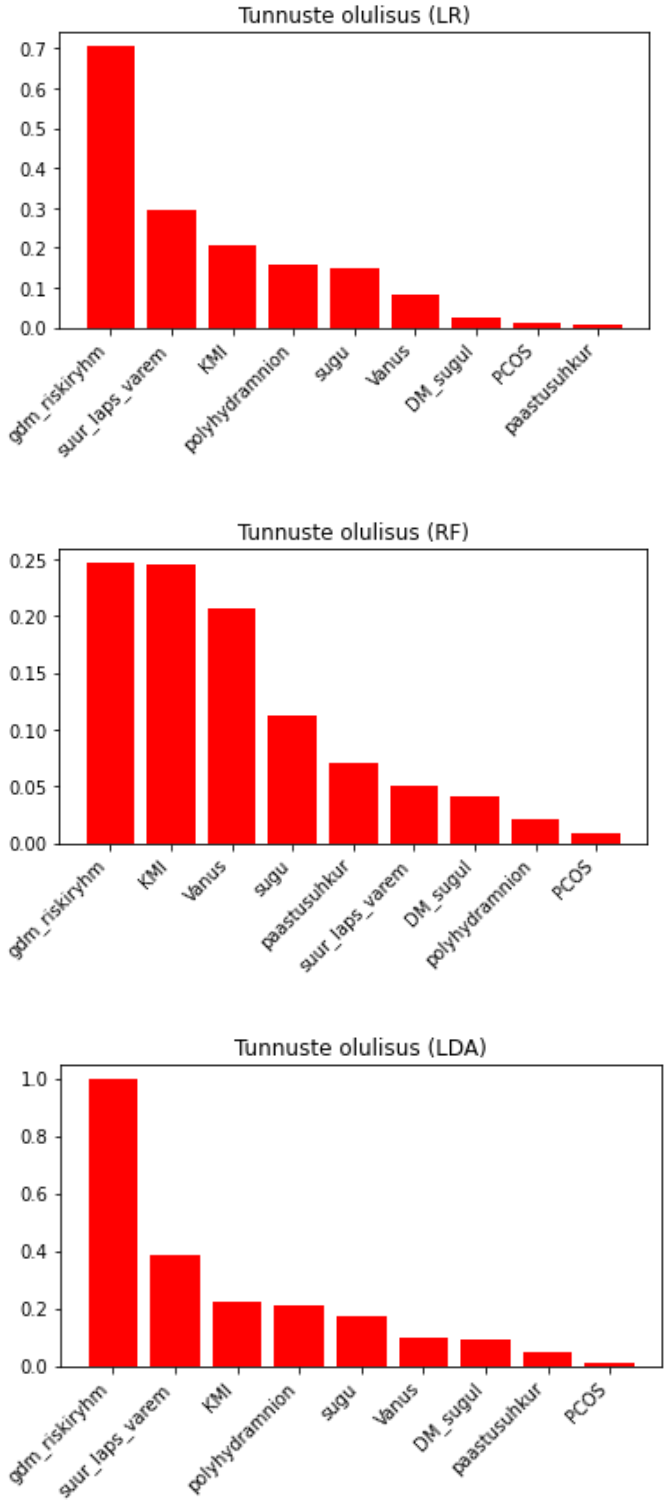


Joonis 10. Vaadeldud ja prognoositud tõenäosuste alusel koostatud kalibreerimisgraafik erinevatele klassifitseerimisalgoritmidele (vaikimisi parameetritega) makrosoomia prognoosimisel. Mudelite lühendid sõnastikus.

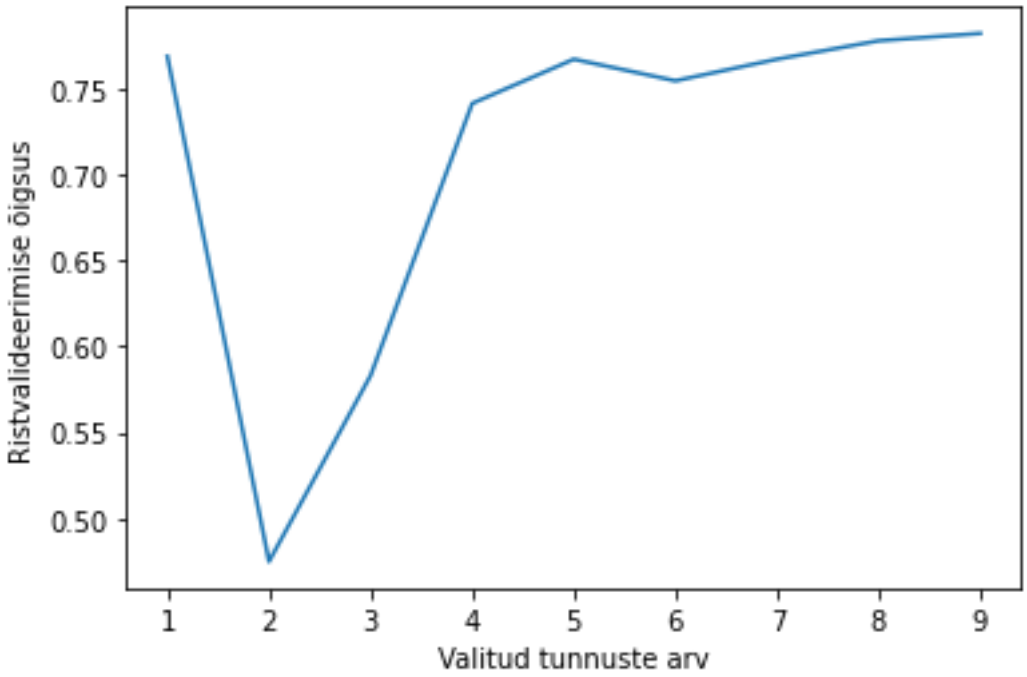
Tabel 7. Makrosoomia ennustusvõime hinnangu kriteeriumid (testandmetel, positiivse klassi kohta) logistilise regressiooni (lineaarne ja polünoomtunnustega), lineaarse diskriminantanalüüsi ja juhusliku metsa mudelite alusel. Kasutusel oli 9 tunnust, hiljemalt raseduse keskel määratavad (vt tabel 6).

Kriteerium/mudel	LR lineaarne	LR polünoomtunnustega	LDA	RF
Valepositiivsete määr (FPR)	0,11	0,01	0,12	0,02
Valenegatiivsete määr (FNR)	0,66	0,43	0,64	0,14
Õigsus (ACC)	0,74	0,87	0,74	0,95
Täpsus (PPV)	0,53	0,95	0,54	0,94
Tundlikkus (TPR)	0,34	0,56	0,36	0,86
Spetsiifilisus (TNR)	0,89	0,99	0,88	0,98
F1	0,41	0,70	0,43	0,90
AUC	0,61	0,77	0,62	0,92

Tunnuste analüüs on toodud joonistel 11 ja 12. Kõikide mudelite jaoks on kõige olulisem tunnus kuuluvus GDM riskirühma (4 rühma: 1) madal risk, ei olnud vaja testida; 2) kõrge risk, kuid ei testitud; 3) kõrge risk, testitud, negatiivne; 4) kõrge risk, testitud, GDM). Sellele järgnevad olenevalt mudelist varasem suure lapse sünd, ema KMI ja ema vanus. Makrosoomia rekursiivse tunnuste eemaldamise puhul (juhusliku metsa mudelis) on optimaalne tunnuste arv 9 ja ühtegi ei saa eemaldada.



Joonis 11. Tunnuste olulisus makrosoomia prognoosimisel (arvutatud mudelite koefitsientide alusel) logistilise regressiooni, juhusliku metsa ja lineaarse diskriminantanalüüsi mudelite puhul. Kasutuses 9 tunnust, mis vaadeldavad hiljemalt raseduse keskpaigas (vt tabel 6)



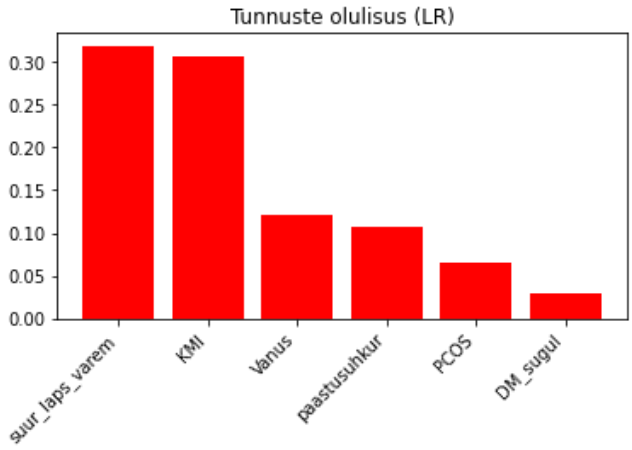
Joonis 12. Tunnuste rekursiivne eemaldamine ristvalideerimise meetodil juhusliku metsa alusel alusel. Õigsus (y-teljel) vastavalt tunnuste arvule (x-teljel). Algselt 9 tunnust (tabel 6).

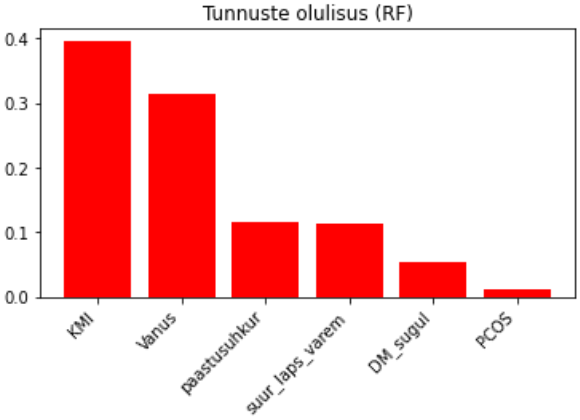
Ka ainult raseduse alguse tunnuste kasutamise puhul teostati analüüs samal viisil. Mudelite valiku jooniseid siin ei näidata, kuid valiti needsamad kolm (LR, LDA, RF). Mudelite ennustusvõime kriteeriumid on ära toodud tabelis 6. Näha on täiendavat ennustusvõime langust võrreldes 9 tunnusega, kuid jälle annab juhuslik mets päris hea tulemuse. Teiste meetodite puhul on päris suur valenegatiivsete määr ja madal tundlikkus, võrreldes 9 tunnusega on langenud ka F1 ja ROC-kõvera alune pindala teistel mudelitel.

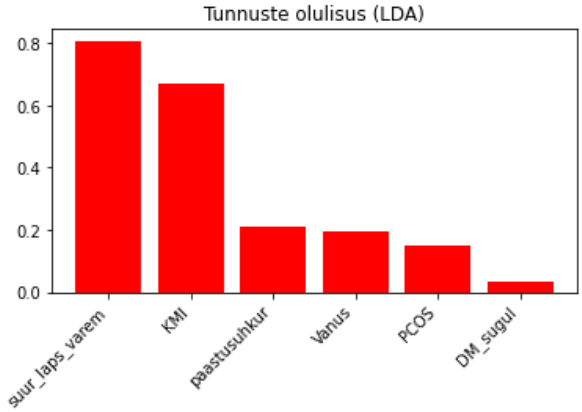
Tabel 8. Makrosoomia prognoosimise hinnangu kriteeriumid (testandmetel, positiivse klassi kohta) logistilise regressiooni (lineaarne ja polünoomtunnustega), lineaarse diskriminantanalüüsi ja juhusliku metsa mudelite jaoks. Kasutusel oli 6 tunnust, raseduse alguses määratavad.

Kriteerium/mudel	LR lineaarne	LR polünoomtunnustega	LDA	RF
Valepositiivsete määr (FPR)	0,01	0,01	0,04	0,01
Valenegatiivsete määr (FNR)	0,88	0,63	0,88	0,17
Õigsus (ACC)	0,82	0,87	0,81	0,96
Täpsus (PPV)	0,83	0,94	0,61	0,94
Tundlikkus (TPR)	0,12	0,37	0,22	0,83
Spetsiifilisus (TNR)	0,99	0,99	0,96	0,99
F1	0,21	0,53	0,32	0,89
AUC	0,56	0,68	0,59	0,91

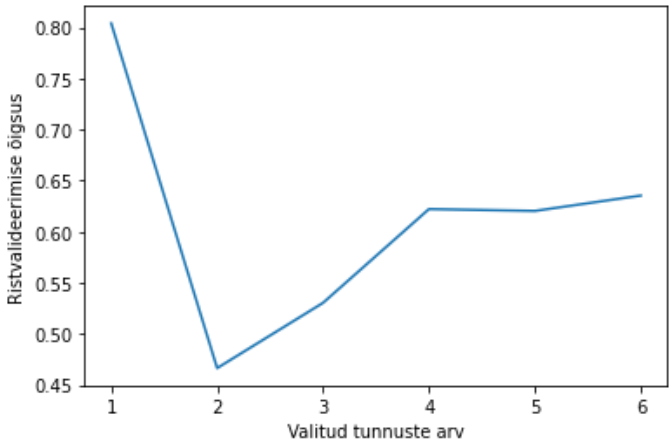
Tunnuste analüüs teostati samuti samal viisil nagu eelnevalt ja tulemused on näha joonistelt 13 ja 14. Kõikide mudelite järgi olid olulisemad ema kehamassiindeks ja suure lapse sünd varasemalt, juhusliku metsa alusel ka ema vanus. Teised tunnused olid vähemtähtsad. Rekursiivse eemaldamise tulemusel juhusliku metsa mudeli puhul jäi alles (oli optimaalne) vaid üks tunnus, ema KMI.







Joonis 13. Tunnuste olulisus makrosoomia prognoosimisel (arvutatud mudelite koefitsientide alusel) logistilise regressiooni, juhusliku metsa ja lineaarse diskriminantanalüüsi mudelite puhul. Kasutuses 6 tunnust, mis vaadeldavad raseduse alguses.



Joonis 14. Tunnuste rekursiivne eemaldamine ristvalideerimise meetodil juhusliku metsa mudeli alusel. Õigsus (y-teljel) vastavalt tunnuste arvule (x-teljel). Algselt 6 tunnust.

6. Järeldused ja arutelu

Gestatsioondiabeet on Eesti andmetel masinõppe meetoditega väga hästi ennustatav. Eriti häid tulemusi andis juhusliku metsa meetod, kuid peaaegu sama hea ennustusvõime oli ka logistilise regressiooni mudelil, kui kasutati polünoomtunnuseid. Sellised tulemused on võrreldes teiste GDM ennustamise töödega üllatavalt head, kuigi käesolevas töös ei kasutatud gradient-võimendatud meetodeid, mis on mitmes uuemas töös head tulemust andnud [45],[47]. Kõik kasutatavad tunnused osutusid ennustamisel vajalikuks ja enamik prognoosimiseks sobivaid tunnuseid on ka määratavad raseduse alguses: ema KMI, ema vanus, PCOS, diabeet lähisugulastel, varasema raseduse ajal laps >4500g, paastusuhkru näit üle normi, ainult polühüdramnion on hilisema tekkega ja lapse sugu hiljem määratav. Kõige olulisemad tunnused olid ema KMI ja paastusuhkru näit, mis mõlemad on juba esimese raseduse alguses määratavad ja seega praktikas hästi rakendatavad. Varasema raseduse ajal GDMi esinemine koondandmestikus puudus, kuid HP andmestiku $\chi 2$-testi põhjal on ka see tunnus oluline ja annaks ennustamisele veelgi juurde, nagu leidsid ka Artzi jt, et varasem GTT tulemus sobib ennustamiseks[45]. Seegi on raseduse alguses teada (kui ei ole esimene rasedus).

GDM-iga rasedatel esineb sagedamini mitmeid raseduse ja sünnitusega seonduvaid probleeme: makrosoomiat, polühüdramnioni, sageli osutub vajalikuks operatiivne sünnitus. Huvitav asjaolu on, et GDM-diagnoosiga rasedatel on kaaluiive oluliselt väiksem kui teistel, see võib näidata, et diagnoos motiveerib neid rasedaid oma toitumisele rohkem tähelepanu pöörama.

Makrosoomia nii hästi meie andmetel ennustatav ei olnud, kuid ka siin olid juhusliku metsa tulemused päris head. Teiste mudelitega läks kõrgeks valenegatiivsete määr ja madalaks tundlikkus. See tähendab, et makrosoomseid lapsi „ei leita üles“. Väga madal oli aga valepositiivsete määr, st „valehäiret“ antakse vähe. Käesoleva töö tulemus on siin mõnevõrra sarnane jaapanlaste tööga [51], kus oli madal täpsus juhusliku metsa mudelit kasutades. Käesolevas töös andis siiski juhuslik mets päris hea täpsuse ja tundlikkuse, küll aga mitte teised mudelid. Makrosoomia puhul oleks siiski olulisem üles leida just võimalikult palju positiivseid ja mõningane valehäire määr oleks isegi rohkem aktsepteeritav.

Kõikide klassifikatsioonide puhul oli kasutatavate andmete jaoks kõige parem juhusliku metsa mudel, üsna häid tulemusi andis ka logistiline regressioon polünoomtunnustega, sarnased tulemused olid ka Akhtar jt töödes [50], [52], [53], kus küll juhuslik mets andis enamasti logistilise regressiooni või SVM järel teise tulemuse. Üldse tunduvad ansamblimeetodid sobivat selle probleemi jaoks paremini [32].

Kui makrosoomia jaoks kasutati ainult riskifaktoreid, mis on määratavad raseduse alguses, siis mudelite ennustusvõime oli veel pisut madalam ja üllatav tulemus oli, et tunnuste rekursiivse eemaldamise tulemusena jäi alles vaid üks tunnus, ema kehamassiindeks enne rasedust. See näitab, et ema ülekaal enne rasedust on üks olulisemaid riskifaktoreid, nii makrosoomia kui ka gestatsioondiabeedi tekkel.

Erinevalt GDMist esines makrosoomiaga lapse sünnituste ajal enam õlgade düstookiat ja perineumi III ja IV astme rebendeid. See võib viidata olukorrale, et makrosoomse lapse sünniks ei olda valmis ja otsustatakse normaalse sünnituse kasuks. GDMi korral ollakse valmis makrosoomse lapse sünniks ja raseduse lõpus hinnatakse oletatavat loote massi ja pigem eelistatakse operatiivset sünnitust.

Makrosoomial on selge seos GDMi ja selle riskirühmaga. Riskirühmadena kasutasime nelja rühma (vt tabel 1). Kõige enam makrosoomseid lapsi sündis rasedate rühmas, kellel oli mingi riskifaktor ja keda testiti GDM suhtes, kuid kes said vastuse, et neil ei ole gestatsioondiabeeti (joonis 3). Võimalik, et põhjuseks on sel juhul rasedal tekkiv tunne, et GDMi ei ole ja järelikult pole ka ohtu liiga suure lapse sünniks ning toitumisele ei pöörata piisavalt tähelepanu. Samas, kui oli diagnoositud GDM, sündis makrosoomseid lapsi vähem, sest need rasedad arvatavasti pöörasid (tõenäoliselt arsti nõuandel) oma toitumisele rohkem tähelepanu. Selline tulemus viitab sellele, et ainuüksi GTT reeglitekohasest rakendamisest ei piisa, vaid ka neil rasedatel, kellel ei ole diagnoositud GDMi, kuid esinevad muud riskitegurid, tuleb rohkem tähelepanu pöörata toitumisele, eriti ülekaalulistel rasedatel.

Kasutatud kirjandus

[1] M. Gillam-Krakauer and C. W. G. Jr, “Birth Asphyxia,” Apr. 2020, Accessed: Jul. 21, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK430782/.

[2] “Bonferroni Correction -- from Wolfram MathWorld.” https://mathworld.wolfram.com/BonferroniCorrection.html (accessed Jul. 30, 2020).

[3] T. Kaart, “Binaarsete tunnuste analüüsimeetodid.” http://ph.emu.ee/~ktanel/bin_tunnuste_analyys/pt26.php (accessed Jul. 21, 2020).

[4] A. Kirss, L. Lauren, M. Rohejärv, and K. Rull, “Gestatsioonidiabeet: riskitegurid, esinemissagedus, perinataalne tulem ja sõeluuringu vastavus juhendile Tartu Ülikooli Kliinikumi naistekliinikus ajavahemikul 01.01.2012–19.06.2013,” *Eesti Arst*, vol. 94, no. 2, pp. 75–82, 2015, doi: 10.15157/ea.v0i0.12060.

[5] R. H. N. Nguyen and A. J. Wilcox, “Terms in reproductive and perinatal epidemiology: 2. Perinatal terms.,” *J. Epidemiol. Community Health*, vol. 59, no. 12, pp. 1019–1021, Dec. 2005, doi: 10.1136/jech.2004.023465.

[6] M. Randväli and V. Kütt, “Muutused rasedusdiabeedi ja selle tüsistuste esinemises seoses uute diagnoosikriteeriumite kasutuselevõtuga,” *Eesti Arst*, vol. 98, pp. 339–343, 2019, Accessed: Jul. 30, 2020. [Online]. Available: https://ojs.utlib.ee/index.php/EA/article/view/15369.

[7] I. Reppo, “Kuidas vältida hüpoglükeemiat,” *Eesti Arst*, Nov. 2016, doi: 10.15157/ea.v0i0.13173.

[8] K. Saks, “Rasvumise paradoks: kehamassiindeksist uute uuringute valguses,” *Eesti Arst*, vol. 97, no. 3, pp. 138–145, 2018, doi: 10.15157/ea.v0i0.14073.

[9] A. Sauga, *Statistika*. 2020.

[10] R. L. Stavis, “Growth Parameters in Neonates - Pediatrics - MSD Manual Professional Edition.” Accessed: Jul. 06, 2020. [Online]. Available: https://www.msdmanuals.com/professional/pediatrics/perinatal-problems/growth-parameters-in-neonates.

[11] M. Tammaru, “KOLMANDA JA NELJANDA JÄRGU LAHKLIHAREBENDITE LEVIMUS JA SEOS EPISIOTOOMIAGA EESTI MEDITSIINILISE SÜNNIREGISTRI ANDMETEL Magistritöö rahvatervishoius,” 2015.

[12] A. Hamza, D. Herr, E. F. Solomayer, and G. Meyberg-Solomayer, “Polyhydramnios: Causes, diagnosis and therapy,” *Geburtshilfe und Frauenheilkunde*, vol. 73, no. 12. Georg Thieme Verlag, pp. 1241–1246, Dec. 2013, doi: 10.1055/s-0033-1360163.

[13] K. Matt and L. Grištšenko, “Polütsüstiliste munasarjade sündroom – kliiniline tähendus,” *Eesti Arst*, Feb. 2004, doi: 10.15157/ea.v0i0.9734.

[14] “Z-Score: Definition, Calculation & Interpretation | Simply Psychology.” https://www.simplypsychology.org/z-score.html (accessed Jul. 25, 2020).

[15] “Mis on andmeteadus? | Data Science Estonia.” Accessed: Apr. 20, 2020. [Online]. Available:

http://datasci.ee/sissejuhatus/mis-on-andmeteadus.

[16] T. M. Mitchell, *The Discipline of Machine Learning*. 2006.

[17] P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York, NY, USA: Cambridge University Press., 2012.

[18] "Avaleht - Masinõpe." Accessed: Mar. 16, 2020. [Online]. Available: https://masinope.ee/.

[19] "Närvivõrkude ja masinõppe sõnastik | Data Science Estonia." Accessed: Mar. 16, 2020. [Online]. Available: http://datasci.ee/masinoppe-sonastik/.

[20] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects.," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[21] R. Raoniar, "Modelling Binary Logistic Regression Using Python (research-oriented modelling and interpretation)," *Towards Data Science*. https://towardsdatascience.com/binary-logistic-regression-using-python-research-oriented-modelling-and-interpretation-49b025f1b510 (accessed Jul. 19, 2020).

[22] R. Gandhi, "Naive Bayes Classifier. What is a classifier?," *Towards Data Science*. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed Jul. 20, 2020).

[23] "Machine Learning Resources | Machine Learning, Deep Learning, and Computer Vision." https://www.ritchieng.com/machine-learning-resources/ (accessed Jul. 31, 2020).

[24] "Masinõppimine - Masinõpe." https://masinope.ee/masinoppimine/ (accessed Jul. 20, 2020).

[25] A. Geron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 2017.

[26] S. Koskel, *Diskriminantanalüüs*. 1998.

[27] R. Agarwal, "The 5 Classification Evaluation metrics every Data Scientist must know," *Towards Data Science*. https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226 (accessed Jul. 23, 2020).

[28] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection *," *Stat. Surv.*, vol. 4, pp. 40–79, 2010, doi: 10.1214/09-SS054.

[29] D. Poulopoulos, "Classifier calibration. The why, when and how of model calibration for classification tasks," *Towards Data Science*. https://towardsdatascience.com/classifier-calibration-7d0be1e05452 (accessed Aug. 02, 2020).

[30] J. Brownlee, *Imbalanced Classification with Python*. 2020.

[31] K. Allen and S. V. F. Wallace, "Fetal macrosomia," *Obstetrics, Gynaecology and Reproductive Medicine*, vol. 23, no. 6. pp. 185–188, Jun. 2013, doi: 10.1016/j.ogrm.2013.03.012.

[32] S. Ye, H. Zhang, F. Shi, J. Guo, S. Wang, and B. Zhang, "Ensemble Learning to Improve the Prediction of Fetal Macrosomia and Large-for-Gestational Age," *J. Clin. Med.*, vol. 9, no. 2, p. 380, Jan. 2020, doi: 10.3390/jcm9020380.

[33] E. Araujo Júnior, A. B. Peixoto, A. C. P. Zamarian, J. Elito Júnior, and G. Tonni, "Macrosomia," *Best Pract. Res. Clin. Obstet. Gynaecol.*, vol. 38, pp. 83–96, Jan. 2017, doi: 10.1016/J.BPOBGYN.2016.08.003.

[34] E. T. Bushman *et al.*, “Influence of Estimated Fetal Weight on Labor Management,” *Am. J. Perinatol.*, vol. 37, no. 3, pp. 252–257, Feb. 2020, doi: 10.1055/s-0039-1695011.

[35] K. Rull, M. Laanpere, and K. Part, “Naistehaigused ja sünnitusabi,” 2010. https://web-proxy.io/proxy/dspace.ut.ee/bitstream/handle/10062/15995/Praktikumid_naistehaigused.pdf?sequence=1&isAllowed=y (accessed Jul. 31, 2020).

[36] D. Roje *et al.*, “Gestational age - The most important factor of neonatal ponderal index,” *Yonsei Med. J.*, vol. 45, no. 2, pp. 273–280, Apr. 2004, doi: 10.3349/ymj.2004.45.2.273.

[37] S. Y. Kim, A. J. Sharma, W. Sappenfield, H. G. Wilson, H. M. Salihu, and O. Gynecol, “Association of Maternal Body Mass Index, Excessive Weight Gain, and Gestational Diabetes Mellitus With Large-for-Gestational-Age Births HHS Public Access Author manuscript,” *Obs. Gynecol*, vol. 123, no. 4, pp. 737–744, 2014, doi: 10.1097/AOG.0000000000000177.

[38] J. Villar *et al.*, “International standards for newborn weight, length, and head circumference by gestational age and sex: The Newborn Cross-Sectional Study of the INTERGROWTH-21st Project,” *Lancet*, vol. 384, no. 9946, pp. 857–868, Sep. 2014, doi: 10.1016/s0140-6736(14)60932-6.

[39] K. Sildver, P. Veerus, and K. Lang, “Sünnikaalukõverad Eestis ja sünnikaalu mõjutavad tegurid: registripõhine uuring,” *Eesti Arst*, vol. 94, no. 8, pp. 465–470, 2015, [Online]. Available: https://eestiarst.ee/sunnikaalukoverad-eestis-ja-sunnikaalu-mojutavad-tegurid-registripohine-uuring/.

[40] K. Kc, S. Shakya, and H. Zhang, “Gestational diabetes mellitus and macrosomia: A literature review,” *Annals of Nutrition and Metabolism*, vol. 66. S. Karger AG, pp. 14–20, Jun. 09, 2015, doi: 10.1159/000371628.

[41] S. Luangkwan *et al.*, “Risk Factors of Small for Gestational Age and Large for Gestational Age at Buriram Hospital,” *J Med Assoc Thai*, vol. 98, 2015, [Online]. Available: http://www.jmatonline.com.

[42] E. Chiefari, B. Arcidiacono, D. Foti, and A. Brunetti, “Gestational diabetes mellitus: an updated overview,” *Journal of Endocrinological Investigation*, vol. 40, no. 9. 2017, doi: 10.1007/s40618-016-0607-5.

[43] A. Holzinger, *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*. Springer Berlin Heidelberg, 2016.

[44] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, and X. Xiao, “Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study,” *J. Diabetes Res.*, vol. 2020, 2020, Accessed: Jul. 24, 2020. [Online]. Available: https://www.hindawi.com/journals/jdr/2020/4168340/.

[45] N. S. Artzi *et al.*, “Prediction of gestational diabetes based on nationwide electronic health records,” *Nature Medicine*, vol. 26, no. 1. Nature Research, pp. 71–76, Jan. 01, 2020, doi: 10.1038/s41591-019-0724-8.

[46] N. Siddegowda Prema and M. P. Pushpalatha, “Analysis of Association between Caesarean Delivery and Gestational Diabetes Mellitus Using Machine Learning,” 2020.

[47] Y.-T. Wu *et al.*, “Early Prediction of High Risk Gestational Diabetes Mellitus via Machine Learning Models,” *SSRN Electron. J.*, Mar. 2020, doi: 10.2139/ssrn.3537076.

[48] H. Qiu *et al.*, “Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, Dec. 2017, doi: 10.1038/s41598-017-16665-y.

[49] E. K. Shriver, “Am I at risk for gestational diabetes? NatioNal iNstitutes of HealtH.”

[50] F. Akhtar *et al.*, “Diagnosis and prediction of Large-for-Gestational-Age fetus using the stacked generalization method,” *Appl. Sci.*, vol. 9, no. 20, Oct. 2019, doi: 10.3390/app9204317.

[51] D. Shigemi, S. Yamaguchi, S. Aso, and H. Yasunaga, “Predictive model for macrosomia using maternal parameters without sonography information,” *J. Matern. Neonatal Med.*, vol. 32, no. 22, pp. 3859–3863, Nov. 2019, doi: 10.1080/14767058.2018.1484090.

[52] F. Akhtar *et al.*, “Diagnosis of large-for-gestational-age infants using a semi-supervised feature learned from expert and data,” *Multimed. Tools Appl.*, pp. 1–31, Jun. 2020, doi: 10.1007/s11042-020-09081-4.

[53] F. Akhtar, “Effective large for gestational age prediction using machine learning techniques with monitoring biochemical indicators,” *J. Supercomput.*, vol. 76, pp. 6219–6237, doi: 10.1007/s11227-018-02738-w.

[54] “Welcome to Python.org.” https://www.python.org/ (accessed Jul. 22, 2020).

[55] “NumPy Reference — NumPy v1.19 Manual.” https://numpy.org/doc/stable/reference/ (accessed Jul. 22, 2020).

[56] “API reference — pandas 1.0.5 documentation.” https://pandas.pydata.org/pandas-docs/stable/reference/index.html (accessed Jul. 22, 2020).

[57] “SciPy — SciPy v1.5.1 Reference Guide.” https://docs.scipy.org/doc/scipy/reference/ (accessed Jul. 22, 2020).

[58] “API Reference — scikit-learn 0.23.1 documentation.” https://scikit-learn.org/stable/modules/classes.html (accessed Jul. 22, 2020).

[59] “API Overview — Matplotlib 3.1.2 documentation.” https://matplotlib.org/3.1.1/api/index.html (accessed Jul. 22, 2020).

[60] “API reference — seaborn 0.10.1 documentation.” https://seaborn.pydata.org/api.html (accessed Jul. 22, 2020).

Lisa 1. Eetikakomitee otsus.

Tartu Ülikooli inimuuringute eetika komitee

Protokolli number: 291/T-3 **koosolek: 18.03.2019**

Komitee koosseis:

Esimees	
Kadri Tamme	Tartu Ülikool, meditsiiniteaduste valdkond, anestesioloogia ja intensiivravi lektor
Aseesimees	
Kristi Lõuk	Tartu Ülikool, humanitaarteaduste ja kunstide valdkond, projektijuht / doktorant
Liikmed	
Diva Eensoo	Tartu Ülikool, meditsiiniteaduste valdkond, analüütik
Naatan Haamer	Tartu Ülikooli Kliinikum, hingehoidja
Ruth Kalda	Tartu Ülikool, meditsiiniteaduste valdkond, peremeditsiini professor / õppetooli juhataja
Malle Kuum	Tartu Ülikool meditsiiniteaduste valdkond, farmakoloogia lektor / farmakoloogia teadur
Liis Leitsalu	Tartu Ülikool, Eesti geenivaramu, genoomika ja geneetilise tagasiside teadur
Maire Peters	Tartu Ülikool, meditsiiniteaduste valdkond, geneetika vanemteadur
Kärt Pormeister	Tartu Ülikool, sotsiaalteaduste valdkond, doktorant
Pille Taba	Tartu Ülikool, meditsiiniteaduste valdkond, neuroloogia professor / kliiniku juhataja
Vahur Ööpik	Tartu Ülikool, meditsiiniteaduste valdkond, spordifüsioloogia professor

Otsus: Kooskõlastada uurimistöö

Uurimistöö nimetus: Gestatsioondiabeedi ja selle riskitegurite levimus ning seos rasedustüsistustega SA TÜK naistekliinikus

Vastutav uurija (asutus):
Kristiina Rull (SA Tartu Ülikooli Kliinikum, naistekliinik, Tartu Ülikool, meditsiiniteaduste valdkond, naistekliinik, Puusepa 8, 50406 Tartu, Eesti)

Komitee poolt läbivaadatud dokumendid:
1. Uurimistöö avaldus kooskõlastuse saamiseks Tartu Ülikool inimuuringute eetika komiteelt, 02.04.2019
2. SA Tartu Ülikooli Kliinikum kooskõlastus uurimistöö toimumise kohta
3. Uurimistöö läbiviijate CVd (K. Rull, A. Kirss, S. Pihu)

Uurimistöö lõpp: 31.12.2020

Komitee esimees: Kadri Tamme */allkirjastatud digitaalselt/*
Komitee sekretär: Kaire Kallak */allkirjastatud digitaalselt/*

Väljastatud: */viimase digitaalallkirja kuupäev/*

Tartu Ülikool
grandikeskus
Lossi 3
51003 Tartu

tel 737 6215
e-post eetikakomitee@ut.ee
www.ut.ee/teadus/eetikakomitee

Lisa 2. Võrkotsingu teel leitud mudelite parameetrid

GDM: **logistiline regressioon** (C=0.012742749857031334, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='saga', tol=0.0001, verbose=0, warm_start=False), **juhuslik mets** (bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=25, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=None, oob_score=False, random_state=1, verbose=0, warm_start=False).

Makrosoomia: **logistiline regressioon** (C=1.623776739188721, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l1', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False), **juhuslik** mets (bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=30, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=300, n_jobs=None, oob_score=False, random_state=1, verbose=0, warm_start=False)

Lisa 3. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Silvia Pihu,**

(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
Gestatsioondiabeedi ja makrosoomia prognoosimine ning nende riskitegurite analüüs masinõppe meetoditega,

(lõputöö pealkiri)

mille juhendajad on Sven Laur ja Kristiina Rull,

(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Silvia Pihu

10.08.2020