TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Jelena Gorbova

Logistilise ja aditiivse logistilise mudeli võrdlus saarlaste antropoloogiliste mõõtmiste näitel

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja dots. Imbi Traat

Tartu 2018

Logistilise ja aditiivse logistilise mudeli võrdlus saarlaste antropoloogiliste mõõtmiste näitel

Lühikokkuvõte:

Bakalaureusetöö eesmärk on võrrelda omavahel logistilist regressioonimudelit ja aditiivset logistilist mudelit binaarses klassifitseerimisülesandes. Töö teoreetilises osas antakse ülevaade klassifitseerimise ideest, logistilisest regressioonimudelist ning põhjalikumalt vaadeldakse üldistatud aditiivset mudelit ja selle erijuhtu – aditiivset logistilist mudelit. Praktilises osas rakendatakse mõlemaid mudeleid saarlaste antropoloogilistel mõõtmistel puht- ja segasaarlaste klassifitseerimiseks. Saadud mudeleid võrreldakse tehisõpe kontekstis, ristvalideerimise teel hinnatakse nende prognoosivõime uute andmete korral.

Võtmesõnad: klassifitseerimine, tehisõpe, prognoosmudelid, antropoloogia

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

Comparision of logistic and additive logistic regression on the example of anthropological measurements of people in Saaremaa

Abstract:

The purpose of the current bachelor thesis is to compare logistic regression and additive logistic regression in the binary classification task. An overview of the classification idea and of the logistic regression is provided in the first chapter. This is followed by more detailed consideration of the generalized additive model and its special case - additive logistic regression. In the second chapter, both models are applied to anthropological measurements to classify pure and mixed inhabitants of Saaremaa. Fitted models are compared with respect to several classification criteria, including the estimated classification accuracy by cross-validation on the new data.

Keywords: classification, statistical learning, predictive models, anthropology

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

Sisukord

1 Sissejuhatus

Logistiline regressioonimudel (logistiline regressioon) on üks levinumaid lähenemisi binaarse klassifitseerimisülesande lahendamiseks. Kuigi tänapäeval eksisteerib mitmeid parameetrilisi ja mitteparameetrilisi meetodeid binaarse tunnuse modelleerimiseks, jääb logistiline regressioonimudel populaarseimaks oma kuju ja interpreteerimise lihtsuse pärast.

Logistilise regressiooni aluseks on lineaarsuse eeldus tõenäosuse $logit$-funktsiooni ja seletavate tunnuste vahel. Reaalses elus on aga andmetes esinevad seosed palju keerulisemad. Mittelineaarseid seoseid lubab aditiivne logistiline regressioon, mis on üldistatud aditiivse mudeli erijuht. Oma kuju poolest on aditiivne logistiline mudel sarnane logistilise regressioonimudeliga, kuid see loobub lineaarsuse nõudest ning lubab rakendada mittelineaarseid funktsioone mudeli seletavatele tunnustele.

Käesoleva bakalaureusetöö eesmärgiks on kirjeldada aditiivset logistilist mudelit ja võrrelda seda klassikalise logistilise regressioonimudeliga. Võrdlus teostatakse tehisõppe kontekstis, kus peamiseks mudeli valideerimiskriteeriumiks on mudeli prognoosivõime uutel andmetel. Tehisõpe on suhteliselt uus statistika haru, mis on pühendatud keeruliste andmekogumite (sealhulgas suurandmete) modelleerimisele ja seaduspärasuste avastamisele. Juurutatud uued meetodid on arvutiintensiivsed ja arendatud sageli paralleelselt nii statistikute kui ka arvutiteadlaste poolt.

Töö esimeses osas vaadeldakse lühidalt logistilise regressiooni kuju ja käsitletakse selle interpreteerimise eripärasid. Tutvustatakse siluvaid splaine kui teatavaid lähendavaid funktsioone. Pikemalt räägitakse üldistatud aditiivsest mudelist ja aditiivsest logistilisest regressioonist. Nii logistilist kui ka aditiivset logistilist mudelit kasutatakse antud töös klassifitseerimiseks. Seetõttu tutvustatakse kriteeriumeid klassifitseerimismudelite valideerimiseks. Lisaks klassikalistele valideerimiskriteeriumitele vaadeldakse ka ristvalideerimise meetodit.

Töö teises osas kirjeldatakse kasutatud andmeid ja rakendatakse neile teoreetilises osas kirjeldatud meetodeid. Mudelite võrdlus teostatakse saarlaste antropoloogiliste mõõtmiste näitel. Nimetatud mõõtmised teostas Eesti antropoloog Juhan Aul 1932. aastal. Käesoleva bakalaureusetöö autor digitaliseeris need unikaalsed, senini üksnes paberkandjal olevad, andmed.

Töö on kirjutatud tekstitöötlusprogrammiga LaTeX. Töös kasutatud R'i koodid on saadaval lisades.

Autor avaldab tänu töö juhendajale Imbi Traadile asjakohaste täienduste, tähelepanelike paranduste ja pühendatud aja eest.

2 Teoreetiline osa

Selles peatükis räägitakse üldiselt klassifitseerimismudeli ideest, vaadeldakse logistilise regressiooni ja üldistatud aditiivse mudeli kuju ning arutletakse klassifitseerimismudeli valideerimiskriteeriume.

2.1 Klassifitseerimine

Antud alapunkti koostamisel kasutati raamatu [15] peatükki 11. Klassifitseerimine on tähtis statistika valdkond, aga viimasel ajal on ta muutunud üheks olulisemaks tehisõppe ülesandeks. See võimaldab teha prognoosi siis, kui uuritav tunnus on mittearvuline ehk kvalitatiivne. Klassifitseerimine on operatsioon, mille käigus omistatakse vaadeldavale objektile selle klassikuuluvus vastavalt tema teatud karakteristikutele.

Olgu $Y = (Y_1, ..., Y_n)^T$ uuritava kvalitatiivse tunnuse juhuslik vektor väärtustega $y = (y_1, ..., y_n)^T$ ja $\mathbf{X}$ seletavate tunnuste $n \times p$-mõõtmeline objekt-tunnus-maatriks:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix},$$

kus n on vaatluste arv andmestikus, p seletavate tunnuste arv ning x_{ij} on i-nda vaatluse j-nda tunnuse väärtus. Tavaliselt esitatakse j-nda tunnuse väärtused vektorkujul:

$$X_j = (x_{1j}, ..., x_{nj})^T.$$

Selle bakalaureusetöö raames vaadeldakse binaarset klassifitseerimisülesannet. Klassifitseerimisülesannet nimetatakse binaarseks siis, kui uuritav tunnus on binaarne ehk võib omada parajasti üht kahest väärtusest. Tuntud binaarsed tunnused on näiteks sugu (mees ja naine) või mingisuguse testi tulemus (edukus ja mitte edukus). Tavaliselt kodeeritakse binaarne tunnus väärtustega 1 ja 0, kus 1-ga tähistatakse huvipakkuva sündmuse toimumist. Kõige levinumad klassifitseerimise meetodid nagu logistiline regressioon või otsustuspuud eeldavad, et andmestikus olevad vaatlused (objektid) on sõltumatud ning uuritava tunnusvektori Y elemendid on Bernoulli jaotusega. Sel juhul kirjeldatakse seos uuritava tunnuse Y ja seletavate tunnuste vahel tingliku tõenäosuse kaudu $\pi_i = P(Y_i = 1|x_{i1}, ..., x_{ip})$, $i \in \{1, ..., n\}$.

2.2 Logistiline regressioonimudel

Logistiline regressioon on üks levinumaid meetodeid binaarse klassifitseerimisülesande lahendamiseks. Logistilise regressioonimudeli konstrueerimisel kasutatakse eelnevalt mainutud lähenemist ja see hindab mitte otseselt Y väärtusi vaid tõenäosust, et vaadeldav objekt kuulub huvipakkuvasse klassi antud seletavate tunnuste väärtuste põhjal.

2.2.1 Logistilise regressioonimudeli kuju ja interpreteerimine

Antud alapunkti koostamisel kasutati raamatut [6] (lk. 119 -127), kui ei ole viidatud teisiti. Logistiline regressioon on üldistatud lineaarse mudeli (ÜLM) erijuht. ÜLM seob eksponentsiaalsete jaotuste perest uuritava tunnuse keskväärtuse seosefunktsiooni seletavate tunnuste lineaarse kombinatsiooniga:

$$g(E(Y|\mathbf{X})) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{1}$$

kus g on seosefunktsioon ning $\beta_j \in \mathbf{R}$, $j \in \{0, ..., p\}$. Logistilise regressiooni puhul on uuritav tunnus Bernoulli jaotusega ning $E(Y|\mathbf{X}) = \pi = (\pi_1, ..., \pi_n)^T$, kus $\pi_i = P(Y_i = 1|x_{i1}, ..., x_{ip})$, $i \in \{1, ..., n\}$. Logistilises regressioonis kasutatakse seosefunktsiooni rollis $logit$-funktsiooni ja selle üldine kuju avaldub järgmiselt:

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{2}$$

kus $\beta_0, ..., \beta_p$ on mudeli hinnatavad parameetrid. Teisisõnu logistiline regressioonimudel seob seletavate tunnuste lineaarset kombinatsiooni huvipakkuva sündmuse tõenäosuse $logit$-funktsiooniga. Paneme tähele, et (2) on võrdus vektorite vahel ja aritmeetilised operatsioonid vektoritele rakendatakse elemendiviisiliselt.

Võrrandist (2) võib avaldada huvipakkuvasse klassi kuuluvuse tõenäosuse (tõenäosuste vektori):

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}. \tag{3}$$

Logistilise mudeli peamiseks eeliseks on selle interpreteerimise lihtsus. Mudeli interpreteerimisel vaadeldakse huvipakkuva sündmuse toimumise shanssi fikseeritud objekti puhul, mis on defineeritud sündmuse toimumise ja mittetoimumise tõenäosuste suhtena [9] (lk. 118):

$$\Pi_i = \frac{\pi_i}{1-\pi_i}, \tag{4}$$

kus $i \in \{1, ..., n\}$.

Kui asendada valemis (4) sündmuse tõenäosus π_i seosest (3), siis sündmuse toimumise shanss taandatakse kujule

$$\Pi_i = e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}.$$

Logistilise regressiooni interpreteerimisel võib tuua esile järgmised seaduspärasused [9] (lk. 118):

- kordaja $\beta_j, j \in \{1, ..., p\}$ positiivne väärtus näitab samasuunalist seost j-nda tunnuse ja sündmuse toimumise shanssi vahel;
- kui vaadelda kahte vaatlust, mille j-nda tunnuse väärtused erinevad ühe ühiku võrra, siis nende shansside suhe on e^{β_j} eeldusel, et teiste tunnuste väärtused nendel vaatlustel ei erine.

2.2.2 Logistilise regressiooni parameetrite hindamine

Logistilise regressiooni parameetrid hinnatakse suurima tõepära meetodiga. Meetodi põhiidee seisneb selles, et leida kordajatele $\beta_0, ..., \beta_p$ sellised hinnangud, et nende asendamisel võrrandisse (3), saadakse ühele lähedane π väärtus nende vaatluste jaoks, mille korral huvipakkuv sündmus toimus. Vastasel juhul peab π olema nullile lähedane [7] (lk. 133). Seda ideed aitab teostada matemaatiline avaldis, tõepärafunktsioon,

$$L(\beta_0, ..., \beta_p) = \prod_{i=1}^{n} \pi_i^{(y_i)}(1 - \pi_i)^{(1-y_i)}, \tag{5}$$

kus y_i omab väärtust 1, kui huvipakkuv sündmus toimus, ja 0 vastasel juhul. Suurima tõepära hinnangud $\hat{\beta}_0, ..., \hat{\beta}_p$ leitakse funktsiooni (5) maksimeerimisel, milleks kasutatakse tõepärafunktsiooni logaritmilist kuju:

$$l(\beta_0, ..., \beta_p) = \log L(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} (y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)).$$

Hinnanguid $\hat{\beta}_0, ..., \hat{\beta}_p$ ei ole võimalik leida analüütiliselt ning tavaliselt kasutatakse selleks iteratiivseid meetodeid, millest kõige levinumad on Newton-Raphsoni ja Fisheri algoritmid [15] (lk. 448).

2.3 Siluvad splainid

Siluvad splainid on kasulikud vahendid mittelineaarsete seoste kirjeldamisel. Olgu $\mathbf{y} = (y_1, ..., y_n)^T$ ja $\mathbf{x} = (x_1, ..., x_n)^T$ vastavalt uuritava ja seletava tunnuste vektorid. Tahetakse leida sellist funktsiooni $g(x)$, mille korral jääkide ruutude summa $RSS = \sum_{i=1}^{n}(y_i - g(x_i))^2$ oleks võimalikult

väike. Suurust $g(x_i)$, $i \in \{1, ..., n\}$ nimetatakse väärtuse y_i prognoosiks. Valides $g(x)$ rolli funktsiooni, mis interpoleerib kõiki $\mathbf{y}$ väärtusi, saadakse alati RSS väärtuseks 0. Kuid sel juhul oleks $g(x)$ liiga paindlik andmete suhtes ning sellise $g(x)$ prognoosivõime uutel andmetel oleks suure kahtluse all [7] (lk. 273-274).

Siluva splaini üldine idee seisneb sellise funktsiooni $g(x)$ leidmises, mis annab nii väikese RSS kui on ka võimalikult sujuv. Nende kahe nõude täitmist tagatakse järgmise funktsiooni minimiseerimisega:

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt, \tag{6}$$

kus $\lambda \in [0, \infty)$ on silumisparameeter. Avaldise esimene pool nõuab, et funktsioon $g(x)$ kirjeldaks $\mathbf{x}$ ja $\mathbf{y}$ seost võimalikult täpselt ning teine pool minimeerib funktsiooni $g(x)$ kurvilisust. Valides sobivat silumisparameetrit λ, leitakse kompromiss nende kahe nõude vahel [6] (lk. 151). Vaatleme kahte piirjuhtu:

- Olgu $\lambda = 0$. Siis (6) teine pool elimineeritakse ning funktsiooniks $g(x)$ võib olla suvaline funktsioon, mis interpoleerib Y väärtusi.
- Olgu nüüd $\lambda = \infty$. Sel juhul $g(x)$ on maksimaalselt sujuv, ehk $g(x)$ on vähimruutude meetodil leitud sirgjoon.

Funktsioon $g(x)$, mis minimiseerib avaldise (6) on naturaalne kuupsplain, kus sõlmede väärtusteks võetakse seletava tunnuse $\mathbf{x}$ kõik unikaalsed väärtused. Naturaalseks kuupsplainiks on funktsioon $S(x)$ [12]:

$$S(x) = \begin{cases} s_1(x), & \xi_1 \leq x < \xi_2, \\ s_2(x), & \xi_2 \leq x < \xi_3, \\ \quad \vdots & \\ s_{K-1}, & \xi_{K-1} \leq x < \xi_K. \end{cases}$$

Funktsioonid s_i on defineeritud järgmiselt:

$$s_i(x) = a_i(x - \xi_i)^3 + b_i(x - \xi_i)^2 + c_i(x - \xi_i) + d_i,$$

kus $i \in \{1, ..., K-1\}$, a_i, b_i, c_i, d_i on reaalarvulised konstandid, ξ_i on sõlmed ja K on sõlmede arv. Siinjuures $S(x)$ peab vastama järgmistele nõuetele:

- funktsioon $S(x)$ läbib kõiki punkte $\xi_1, ..., \xi_K$;

- funktsioon $S(x)$, selle esimene ja teine tuletis on pidevad igas punktis $\xi_1, ..., \xi_K$;
- $S(x)$ on lineaarne väljaspool $[\xi_1, \xi_K]$.

Paneme tähele, et silumisparameeter λ kontrollib mitte ainult funktsiooni $g(x)$ kurvilisust, vaid ka vabadusastmete arvu. Võib näidata, et siluva splaini puhul vabadusastmete arv muutub n-st 2-ni ning mida kõrgem on vabadusastmete arv, seda paindlikum on $g(x)$. Silumisparameetri λ ja vabadusastmete arvu vahel esineb vastassuunaline seos, mis tähendab, et λ suurendamisel kaheneb funktsiooni vabadusastmete arv [6] (lk. 153-156).

2.4 Üldistatud aditiivne mudel

Antud alapunkt põhineb monograafiale [5], kus Hastie ja Tibshirani tõid sisse üldistatud aditiivse mudeli (ÜAM). ÜAM üldine kuju avaldub järgmiselt:

$$g(E(Y|\mathbf{X})) = \alpha + f_1(X_1) + ... + f_p(X_p). \tag{6}$$

Siin α on vabaliige ning $f_j, j \in \{1, ..., p\}$ võib olla nii parameetriline kui ka mitteparameetriline funktsioon.

Üldistatud aditiivne mudel seob uuritava tunnuse keskväärtuse seosefunktsiooni seletavate tunnuste kombinatsiooniga, kus igale tunnusele rakendatakse ühemuutuja funktsiooni. Summeerimise tõttu mudel kannabki nime aditiivne. Kuigi mudeli aditiivne kuju paneb loobuma koosmõjude arvestamisest, lihtsustab see mudeli interpreteerimist. Sarnaselt ÜLM-ga lubab üldistatud aditiivse mudeli kuju hinnata j-nda tunnuse mõju uuritavale tunnusele fikseerides ülejäänud seletavad tunnused. Võrreldes ÜLM-ga on üldistatud aditiivsel mudelil üks suur eelis. ÜAM võimaldab rakendada seletavatele tunnustele funktsioone, mis võivad igal tunnusel olla erinevad, nii parameetrilised kui ka mitteparameetrilised. Seega üldistatud aditiivne mudel on paindlikum kui lineaarne üldistatud mudel, sest lubab arvestada mitte ainult lineaarsete, vaid ka mittelineaarsete seostega.

2.4.1 Aditiivne mudel ja selle hindamine

Üldistatud aditiivse mudeli lihtsaimaks erijuhuks on aditiivne mudel:

$$E(Y|\mathbf{X}) = \alpha + f_1(X_1) + ... + f_p(X_p). \tag{7}$$

Seega tavaline aditiivne mudel hindab otseselt uuritava tunnuse Y keskväärtust. Funktsioonide $f_1, ..., f_p$ hinnangud leitakse lähtudes jääkide ruutude summa modifikatsioonist:

$$\sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}))^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j, \tag{8}$$

kus $\lambda_j \geq 0$ on silumisparameeter. Avaldise (8) minimiseerimisel konstrueeritakse iteratiivselt iga j jaoks siluv splain, kusjuures f_j algväärtuste määramiseks eeldatakse, et iga j korral $\sum_{i=1}^{n} f_j(x_{ij}) = 0$. Sel juhul on mudeli vabaliikme hinnanguks $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Funktsioonide $f_1, ..., f_p$ hinnangute leidmise käik on toodud Algoritmis 1, mis on tuntud inglisekeelses kirjanduses *backfitting* nime all.

Algoritm 1: *Backfitting* algoritm aditiivse mudeli hindamiseks.

1. Arvutame: $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\hat{f}_j \equiv 0 \; \forall j \in \{1, ..., p\}$.

2. Kordame kuni $\hat{f}_j$ koondub:

$$\hat{f}_j = S_j \left[(y - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(X_j)) \right],$$

kus S_j on siluva splaini funktsioon.

2.4.2 Aditiivne logistiline mudel

Sarnaselt ÜLM-ga võib üldistatud aditiivse mudeli puhul uuritav tunnus olla nii arvuline kui ka kvalitatiivne. Binaarse uuritava tunnuse modelleerimiseks kasutatakse tavaliselt aditiivset logistilist mudelit.

Aditiivne logistiline mudel on üldistatud aditiivse mudeli erijuht, kus seosefunktsiooni rollis kasutatakse $logit$-funktsiooni. Aditiivse logistilise mudeli kuju avaldub järgmiselt:

$$\log \frac{\pi}{1 - \pi} = \alpha + f_1(X_1) + ... + f_p(X_p), \tag{9}$$

kus $\pi = (\pi_1, ..., \pi_n)^T$, $\pi_i = P(Y_i = 1 | x_{i1}, .., x_{ip})$.

Sarnaselt tavalise logistilise regressiooniga hinnatakse aditiivne mudel tõepärafunktsiooni maksimeerimisel. Hinnangute $\hat{f}_1, ..., \hat{f}_p$ leidmiseks kasutatakse lokaalse skooringu protseduuri, mis on esitatud Algoritmis 2. Lokaalse skooringu protseduur on kahekordne iteratiivne meetod üldistatud aditiivse mudeli hindamiseks, kus välimises tsüklis rakendatakse Newton-Raphsoni

algoritmi tõepärafunktsiooni maksimeerimiseks ning sisemises pöördutakse eelnevalt kirjeldatud *backfitting* algoritmi poole.

Algoritm 2: Lokaalse skooringu algoritm aditiivse logistilise mudeli hindamiseks

1. Arvutame algväärtuse $\hat{\alpha} = \log \frac{\bar{y}}{1-\bar{y}}$, kus $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, ja $\hat{f}_j \equiv 0, \forall j \in \{1, ..., p\}$.

2. Defineerime abimuutujad $\hat{\mu}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ ja $\hat{p}_i = \frac{1}{1+e^{-\hat{\mu}_i}}$, kus indeks i määrab vaatluse (objekti) järjekorra numbri.

   Moodustame kaalud w_i ja ajutise uuritava tunnuse z_i:

   $$w_i = \hat{p}_i(1 - \hat{p}_i),$$

   $$z_i = \hat{\mu}_i + \frac{y_i - \hat{p}_i}{w_i}.$$

   Hindame kaalutud aditiivse mudeli z_i suhtes. Saame uued hinnangud $\hat{f}_j \; \forall j$.

3. Kordame punkti 2, kuni $\hat{f}_j, \forall j$ koondub.

2.5 Klassifitseerimismudeli headuse näitajad

2.5.1 Akaike informatsioonikriteerium

Akaike informatsioonikriteerium AIC [1] (lk. 261-281) on üks levinuim mudeli headuse näitaja, mis võimaldab võrrelda omavahel seotud mudeleid. AIC on defineeritud järgmiselt:

$$2K - 2 \log L(\Theta),$$

kus K on hinnatavate parameetrite arv mudelis, $L(\Theta)$ on maksimeeritud tõepära funktsioon ja Θ on hinnatud parameetrite vektor. AIC kordaja on lihtne ja efektiivne vahend mudeli headuse hindamiseks, mis arvestab mitte ainult prognoosi täpsusega, vaid ka mudeli keerukusega. Parimaks mudeliks peetakse mudelit, mille AIC väärtus on minimaalne, kuid AIC väärtust iseennast ei ole võimalik interpreteerida.

Mitteparameetriliste mudelite puhul arvutatakse Akaike informatsioonikriteeriumi modifikatsioone, mis kasutavad tinglikku või marginaalset tõepärafunktsiooni. Hastie ja Tibshirani [5] defineerisid AIC üldistatud aditiivsete mudelite jaoks järgmiselt:

$$\text{AIC} = D(y, \hat{y})/n + 2df\phi/n,$$

kus $D(y, \hat{y})$ on mudeli hälbimus, mis aditiivse mudeli puhul leitakse asümptootiliselt, n on vaatluste arv, df vabadusastmete arv ja ϕ dispersiooni parameeter [5] (lk. 158).

2.5.2 Klassifitseerimisviga ja seotud kriteeriumid

Klassifitseerimise tulemuslikkuse iseloomustamisel kasutatakse peatükki 2.2 raamatus [11]. Olgu tegemist binaarse klassifitseerimismudeliga, millega prognoositakse objekti klassi. Tähistame prognoositud y väärtused tähega $\hat{y}$. Kuna tegemist on binaarse uuritava tunnusega, siis võib väärtustest $\hat{y}$ ja y moodustada järgmise 2×2 sagedustabeli, kus sümbol # tähistab loendustehet:

Õige positiivsus (ÕP) #($\hat{y} = 1$ ja $y = 1$)	Vale positiivsus (VP) #($\hat{y} = 1$ ja $y = 0$)
Vale negatiivsus (VN) #($\hat{y} = 0$ ja $y = 1$)	Õige negatiivsus (ÕN) #($\hat{y} = 0$ ja $y = 0$)

Siin õige positiivsus on objektide arv, mis tegelikult kuuluvad klassi 1 ja mida õigesti klassifitseeriti klassi 1. Vale positiivsus on objektide arv, mille tegelik klassikuuluvus on 0, kuid klassifitseerimisel määrati nad klassi 1. Nagu on näha, moodustavad klassifitseerimistabeli peadiagonaalil olevad sagedused kokku õigesti klassifitseeritud objektide arvu. Kasutades ülalkirjeldatud sagedusi võib defineerida olulised klassifitseerimismudeli tulemuslikkuse näitajad:

Täpsus	Klassifitseerimisviga	Tundlikus	Spetsiifilisus
$\frac{\text{ÕP+ÕN}}{\text{ÕP+VP+VN+ÕN}}$	$\frac{\text{VP+VN}}{\text{ÕP+VP+VN+ÕN}}$	$\frac{\text{ÕP}}{\text{ÕP+VN}}$	$\frac{\text{ÕN}}{\text{ÕN+VP}}$

Klassifitseerimismudeli täpsus ja klassifitseerimisviga on teineteist täiendavad näitajad. See tähendab, et teades üht näitajat on lihtne tuletada teist, mistõttu praktikas valitakse tavaliselt neist ainult üks.

2.5.3 Ristvalideerimine

Vaadeldav alapunkt põhineb 5. peatükile raamatus [7]. Ristvalideerimine on statistilise mudeli valideerimismeetod, mille abil on võimalik hinnata mudeli prognoosi täpsust uutel andmetel.

Statistilise mudeli valideerimiseks jagatakse andmestik tavaliselt kaheks mittelõikuvaks osaks: treeningandmeteks ja test- ehk valideerimisandmeteks. Mudel konstrueeritakse kasutades

treeningandmeid. Seejärel tehakse prognoos valideerimisandmetele kasutades eelnevalt konstrueeritud mudelit ning arvutatakse testviga, mis on prognoosivea hinnanguks. See lähenemine võib anda suure varieeruvusega prognoosivea hinnangu, kuna see sõltub treening- ja valideerimisandmete jaotusest. Samuti ei ole seda meetodit mõistlik kasutada väikeste andmestike puhul. Jagades andmeid kaheks osaks loobutakse mudeli treenimisel suurest osast andmetest, mis kindlasti mõjutab prognoosimise tulemuslikkust.

Ristvalideerimise meetod elimineerib need puudused. Tänapäeval kasutatakse mitut ristvalideerimise modifikatsiooni. Selles bakalaureusetöös vaadeldakse K-kordset ristvalideerimise meetodit, kuna see annab optimaalselt väikese nihkega ja varieeruvusega testvea.

K-kordsel ristvalideerimisel jagatakse andmestik juhuslikult K võrdseks osavalimiks. Üks osavalimitest valitakse valideerimisandmeteks ning ülejäänuid $K-1$ osavalimit kasutatakse mudeli treenimiseks. Klassifitseerimismudeli puhul arvutatakse igal iteratsioonil treenitud mudeli klassifitseerimisviga:

$$KV_k = \frac{\#(y_k \text{ != } \hat{y}_k)}{N_k},$$

kus y_k on k-nda iteratsiooni valideerimisandmete uuritava tunnuse väärtused, $\hat{y}_k$ vastavad prognoositud väärtused, N_k vaatluste arv k-ndas osavalimis, $k \in \{1, ..., K\}$. K-kordsel ristvalideerimisel kasutatakse iga osavalimit valideerimisandmetena parajasti üks kord. Viimasel sammul arvutatakse keskmine mudeli täpsuse näitaja üle K korduse ning ristvalideerimise prognoosivea hinnang avaldub järgmiselt:

$$RV = \frac{\sum_{k=1}^{K} KV_k}{K}.$$

3 Praktiline osa

Töö praktilises osas võrreldakse logistilist mudelit ja aditiivset logistilist mudelit binaarses klassifitseerimisülesandes. Võrdlus teostatakse Juhan Auli poolt kogutud saarlaste andmete näitel. Järgmistes alapunktides räägitakse kasutatud andmete päritolust, kirjeldatakse tunnuste valimise protsessi, rakendatakse teoreetilises osas kirjeldatud mudeleid ja analüüsitakse tulemusi. Andmete töötlemiseks ja mudelite võrdlemiseks on kasutatud tarkvara R. Kasutatud R'i koodid on toodud Lisas 2.

3.1 Andmete kirjeldus

Professor Juhan Aul on Eesti antropoloog ja zooloog. Ta oli üks esimestest teadlastest, kes on tõsiselt uurinud eestlaste antropoloogilist kuuluvust [4]. Oma teadustöödes uuris ta nii kogu Eesti ala kui ka selle väiksemate piirkondade rahvastiku antropoloogiat ning tegi võrdlusanalüüsi naaberriikide rahvastega. Aastal 1933 ilmus tema raamat "Maailmasõja antropoloogilisest mõjust saarlastele"[3], kus ta uuris Eesti puht- ja segasaarlaste antropoloogilisi erinevusi. Idee uurimuseks andis talle vallaslaste osakaalu järsk tõus Saaremaa lõuna-, lääne- ja loodepiirkondades esimese maailmasõja ajal aastatel 1916-1918 (Tabel 3.1). Selle nähtuse võimalikuks põhjuseks pakkus ta esimese maailmasõja ajal Saaremaal ja selle ümbruses toimunud sündmusi.

Esimese maailmasõja ajal olid Muhu väin ja Läänemere saared pinevuskoldeks Läänemere piirkonnas. Eesti kuulus siis Tsaari-Venemaa koosseisu. Saaremaa ja teised ümberkaudsed saared said oma soodsa asendi, võimsa kaitsesüsteemi loomise võimaluste ja küllaldaste laevastikubaaside tõttu strateegiliseks sihtmärgiks Vene ja Saksa vägede jaoks [14]. Aastal 1914 jõudis sõda Eesti pinnale koos Saksa vägedega. Juba aastaks 1915 näitasid sakslased oma kavatsuste tõsidust teostades dessandi Ruhnu saarel. Hiljem tulistati ka Sõrve ja Roomassaare sadamaid. Samal ajal kindlustasid Vene väed omalt poolt Saaremaad ja teisi lähedasi saari, kuna Vene sõjaväe juhtkond uskus, et need etendavad impeeriumi pealinna kaitsel olulist osa [10]. Loodi Muhu Väina Kindlustatud Positsioon (MVKP), mille paiknemiskohti aastal 1915 võib näha Joonisel 4.4 (Lisa 1). Aastaks 1917 tekkis Muhu väina piirkonnas ja mujal Saaremaal üle saja sõjalise objekti, mille relvastuses olid igasugused rannakaitsesuurtükid ning õhutõrjekahurid. Rannakaitsekahureid teenindas kokku 1500 mereväelast, kes suuremas osas olid venelased [13].

Juhan Aul vestles 1932. aastal laste emade, õpetajate ja vanemate tegelastega ning selgitas välja, et tõepoolest suuremas osas olid sõja aastatel sündinud vallaslaste isad vene sõjaväelised, kes tulid sinna piirkonda sõjategevuses osalemiseks [3]. Antropoloogil Aulil tekkis huvi, kas

Tabel 3.1. Jämaja, Anseküla, Kihelkonna, Mustjala koguduste vallaslaste jaotus 1908-1922 [3].

Aasta	**Sündinute arv**	**Nendest vallaslapsi**	**Vallaslapsed (protsentides)**
1908	465	26	5.6
1909	469	33	7.0
1910	503	37	7.4
1911	487	32	6.6
1912	483	49	10.1
1913	422	41	9.7
1914	486	48	9.9
1915	420	44	10.5
1916	**287**	**58**	**20.2**
1917	**311**	**104**	**33.4**
1918	**351**	**84**	**23.9**
1919	**347**	**37**	**10.7**
1920	304	37	12.2
1921	402	35	8.7
1922	415	47	11.3

ja millised somaatilised erinevused esinevad lapsel, kelle ema on pärit Saaremaalt, aga isa on muulane, võrreldes puhtsaarlastega. Tema antropoloogilise analüüsi aluseks said 13- kuni 16-aastaste laste üldfüsioloogilised mõõtmised, mida ta teostas Saaremaal 1932. aasta kevadel. Näeme, et uuringus osalevate laste sünniaastad varieeruvad 1916st kuni 1919ni. Tabeli 3.1 kohaselt toimus mainitud aastatel sündinud vallaslaste osakaalus suur tõus, mis põhjendab Auli valikut katseisikute vanuse osas.

Kokku osales Auli uuringus 411 last, kellelt võeti järgmised mõõdud: pea suurim pikkus, pea suurim laius, lauba vähim laius, näolaius, lõualaius, pea kõrvakõrgus, pea üldkõrgus, füsiognoomiline ja morfoloogiline näokõrgus, nina kõrgus ja laius, üldpikkus, siruulatus, õlakõrgus, sõrmekõrgus, rinnakukõrgus, sümfüüsi kõrgus, iliospinaal, istepikkus, õlalaius, rinnalaius, rinnasügavus, puusalaius, puusa eesnurga kõrgus, pea ümbermõõt, rinna ümbermõõt, talje ümbermõõt, õlavarre, käsivarre, reie ümbermõõt, sääre pikkus, kurgu pikkus, raskus, käe, jala ja kere pikkus. Kõik mõõtmistulemused kanti ankeedilehele (Joonised 4.1, 4.2, Lisa 1) mil-

limeetrites. Lisaks antropoloogilistele mõõtmistele fikseeriti iga lapse kohta, mis rahvusest on tema vanemad, vanemate ja lapse päritoluvald, vanemate elukutse (valikvastustega talunik, väiketalunik, vabadik, asunik, põllutööline, vabrikutööline, ametnik, käsitööline, õpetaja, kaupmees), õdede ja vendade arv ning mitmenda lapsena küsitletav ise on sündinud. Samuti kanti ankeeti ema ja isa vanus, lapse juuste värv (valikvastustega must, pruun, blond, ruuge, hele, tume, kollakas, tuhkjas) ja silmade värv (valikvastustega pruun, tumekirju, sinine, hall, hele, tume, rohekas, kollakas, valkjas).

Nagu juba eelpool mainitud olid Juhan Auli poolt kogutud andmed kantud ankeetidesse, mis asuvad hetkel Tartu Ülikooli muuseumi arhiivis [2]. Andmed olid senini saadaval ainult paberkujul. Selle bakalaureusetöö autor digitaliseeris need andmed. Tulemuseks on 57 tunnusega andmestik, mis sisaldab infot 411 lapse kohta. Lisaks eelnevalt nimetatud tunnustele sisaldab saadud andmestik vallaslapseks olemise tunnuse. Last vaadeldi vallaslapsena või mitte sõltuvalt sellest, kas ankeedis lapse nime kõrval oli kirjas ema või isa nimi. Näiteks Joonisel 4.2 (Lisa 1) toodud ankeedi puhul on lapse nime kõrval kirjas ”Anne t.”, ehk tegemist on vallaslapsega. Juhul kui ankeedis polnud infot ema ega isa kohta, siis tunnuse **Vallaslaps** väärtuseks sai ”ei ole teada”.

Antud bakalaureusetöös võeti uuritavaks tunnuseks tunnus **Saarlane** tasemetega ”Puhtsaarlane” ja ”Segasaarlane”. Laps oli puhtsaarlane siis, kui mõlemad tema vanemad olid eestlased. Uurides lähemalt lapse vanemate rahvust, selgus, et andmestikus esines 25 last, kelle ema või isa oli saksa, soome, tatari, poola, leedu või muust rahvusest. Kuna nimetatud 25 lapse osakaal Auli poolt mõõdetud laste hulgas on ainult 6%, siis arvestades ka eelnevat arutelu vallaslaste sündide tõusu ja vene sõjavägede asukoha seose kohta, otsustati järgnevasse analüüsi võtta ainult eesti ja vene isadega lapsed. Kokku jäi analüüsi 386 last: 87 last kelle isa on venelane ja ema on eestlane (segasaarlane) ja 299 last, kelle mõlemad vanemad on eestlased (puhtsaarlane).

3.2 Andmete esmane töötlus ja tunnuste valimine

Andmete digitaliseerimisel selgus, et mõõtmistega on tegelenud mitu erinevat inimest, mistõttu kohati muutub tunnuste komplekt ja sisestamise viis. See asjaolu tekitab raskusi mõnede tunnuste interpreteerimisel. Näiteks lapse vanemate elukutse puhul on paljudel juhtudel valitud ainult üks elukutse, kuid mõnel juhul on märgistatud kaks elukutset ilma täpsustamata, kumb on ema ja kumb isa oma.

Selles alapunktis käsitletakse puuduvate väärtuste probleemi, kodeeritakse mõned olulised

tunnused ümber ning valitakse tunnused järgneva analüüsi jaoks.

3.2.1 Tunnuste ümberkodeerimine

Auli ankeetides oli esitatud mitu valikut silmade ja juuste värvi kirjeldamiseks, täpsemalt juuste värv (must, pruun, blond, ruuge), juuste värvi toon (hele, tume, kollakas, tuhkjas), silmade värv (pruun, tumekirju, sinine, hall) ning silmade värvi toon (hele, tume, rohekas, kollakas, valkjas). Need kvalitatiivsed tunnused sisaldavad väga spetsiifilist informatsiooni. Selleks, et kirjeldada silmade ja juuste värvi üldisemalt moodustati kaks uut tunnust järgmiste tasemetega:

- **Juuksed**: blond, pruun, tume, ruuge, muu;
- **Silmad** sinine, kirju, hallikas, tume kirju, pruun, muu.

Kolmas mitme tasemega kvalitatiivne tunnus on **Päritolu**. Andmete digitaliseerimisel lapse **Päritolu** märgistati valla tasemel. Tunnusel **Päritolu** on kokku 15 taset: Abruka, Kaarma, Kärla, Kihelkonna, Kuressaare, Leisi, Lümmada (Lümanda), Maasi, Muhu, Mustjala, Pihtla, Torgu ja Uuemõisa. Valdade asukohad on kujundatud Joonisel 4.3 (Lisa 1). Arvestades Saaremal toimunud sõjategevusega, mida kirjeldati eelmises peatükis, otsustati tunnuse **Päritolu** põhjal moodustada uus tunnus, mis näitaks, kas lapse sünnivallas olid aastatel 1915-1917 Vene väed või mitte. Tunnuse moodustamisel kasutati rannakaitsesuurtükkide positsioonide kaarte aastatel 1915 ja 1917 (Joonised 4.4, 4.5, Lisa 1). Uue tunnuse **Päritolu2** väärtuste geograafiline jaotus on esitatud Joonisel 4.3 (Lisa 1).

3.2.2 Puuduvad väärtused

Andmestiku digitaliseerimisel selgus, et paljudel tunnustel on suur puuduvate väärtuste osakaal. Otsustati kasutada ainult neid tunnuseid, milles puudutavate väärtuste osakaal ei ületa 10%. Selliseid tunnuseid on andmestikus 33, vastavalt andmestikus kasutatud nimedega: Sugu, Saarlane, Vanus, Pea pikkus, Pea laius, Lauba laius, Näo laius, Pea kõrvakõrgus, Pea üldkõrgus, F. näokõrgus, Kere pikkus, M. näokõrg, Juuksed, Silmad, Päritolu2, Lõua laius, Üldpikkus, Õlalaius, Vanemate elukutse, Õlakõrgus, Sõrmekõrgus, Iliospinaal, Jala pikkus, Istepikkus, Käe pikkus, Vallaslaps, Suprastern, Sümfüüs, Rinna ümbermõõt, Puusalaius, Vennad, Õed ja tunnus Mitmes.

Tunnus **Vanemate elukutse** otsustati andmestikust välja võtta seoses andmete ebakorrektse sisestamisega. Selleks, et vältida multikollineaarsust võeti välja tunnused **Vennad**, **Õed**, kuna need tunnused on tugevalt korreleeritud tunnusega **Mitmes**.

Tabel 3.2. Valitud tunnuste loetelu ja kirjeldus. Veerus NA on puuduvate väärtuste arv.

Tunnus	**Tunnuse kirjeldus**	**NA**
Saarlane	binaarne tunnus, puht-/segasaarlane	0
Sugu	binaarne tunnus, poiss/tüdruk	0
Vanus	lapse vanus aastates	0
Pea pikkus	pidev tunnus	0
Pea laius	pidev tunnus	0
Pea kõrvakõrgus	pidev tunnus	0
Pea üldkõrgus	pidev tunnus	0
M. näokõrgus	morfoloogiline näokõrgus, pidev tunnus	0
Juuksed	juuste värv, blond/pruun/tume/ruuge/muu	0
Silmad	silmade värv, sinine/kirju/hallikas/tume kirju/pruun/muu	0
Päritolu2	binaarne, VeneJah/VeneEi (Vene vägede asukoht)	0
Vallaslaps	jah/ei/ei ole teada	0
Lauba laius	lauba vähim laius, pidev tunnus	1
Näo laius	põsekaarte vahe, pidev tunnus	1
F. näokõrgus	füsiognoomiline näokõrgus, pidev tunnus	1
Lõua laius	lõuanurkade vahe, pidev tunnus	2
Üldpikkus	pidev tunnus	2
Õlalaius	pidev tunnus	2
Õlakõrgus	pidev tunnus	6
Sõrmekõrgus	pidev tunnus	6
Iliospinaal	iliospinaalkõrgus, pidev tunnus	7
Jala pikkus	pidev tunnus	9
Käe pikkus	õlanuki ja keskmise sõrme vaheline kaugus, pidev tunnus	10
Istepikkus	keha pikkus istudes, pidev tunnus	10
Rinna ümbermõõt	pidev tunnus	24
Suprastern	kaela pikkus, pidev tunnus	26
Sümfüüs	süleliiduse kõrgus, pidev tunnus	26
Kere pikkus	rinnaku- ja süleliidusekõrguse vahe, pidev tunnus	27
Puusalaius	pidev tunnus	29
Mitmes	mitmendana laps on sündinud peres, diskreetne tunnus	29

Tabelis 3.2 on esitatud valitud tunnused koos kirjelduse ja puuduvate väärtuste arvuga. Üheks levinuks puuduvate andmete käsitlusmeetodiks on puuduvate väärtuste asendamine vaadeldava tunnuse keskväärtusega [9]. Antud töös rakendatakse seda käsitlusmeetodit kõikidele pidevatele tunnustele. Kuna diskreetse tunnuse **Mitmes** puuduvate väärtuste osakaal on ainult 8 % ning neile ei ole mingit loogilist asendust, võeti 29 vaatlust selle tunnuse puuduva väärtusega andmestikust välja.

3.3 Logistilise regressiooni rakendamine

Esmalt konstrueeritakse logistiline regressioonimudel puht- ja segasaarlaste klassifitseerimiseks. Tehisõppe kontekstis pööratakse mudeli valideerimisel tähelepanu mitte ainult sellele, kuidas mudel kirjeldab käesolevaid andmeid, vaid ka mudeli prognoosivõimele uute objektide (siin laste) klassifitseerimisel [7].

Pärast andmete esmast töötlemist jäi andmestikku 357 vaatlust ja 30 tunnust. Kuna tegemist on suhteliselt väikese andmestikuga, siis mudeli prognoositäpsuse hindamiseks ei jagata andmestikku treening- ja valideerimisandmeteks. Tunnuste komplekti valimisel ja mudeli headuse näitajate arvutamisel kasutatakse kogu andmestikku treeningandmetena. Testvea arvutamiseks kasutatakse aga K-kordset ristvalideerimise meetodit.

Esimesel sammul kasutati mudeli konstrueerimiseks kõiki 30 tunnust. Enamus tunnuseid osutusid ebaolulisteks. Parima tunnuste komplekti valimiseks kasutati sammregressiooni. Sammregressiooni teostas tarkvara R funktsioon $stepAIC$, mis on defineeritud paketis MASS. Tuginedes AIC kriteeriumile pakutakse tulemuseks mudel, mille AIC kriteeriumi väärtus on minimaalne üle vaadeldud tunnuste kombinatsioonide. Kasutades nii ettepoolset kui ka tahapoolset tunnuste valiku meetodit saadi tulemuseks üks ja sama mudel, mis kasutab 9 tunnust AIC kordaja väärtusega 198. Minimaalse AIC väärtuse korral osutus tunnus **Pea üldkõrgus** ($p = 0.14$) ebaoluliseks olulisuse nivool 0.1. Eemaldades mudelist selle tunnuse kasvas AIC kordaja väärtus ainult 0.23 võrra ehk tunnuse **Pea üldkõrgus** eemaldamine mudelist eriti ei mõjutanud AIC kordajat. Samuti näitas Hii-ruut test olulise nivool 0.1 ($p = 0.14$), et mudel, mis sisaldab tunnust **Pea üldkõrgus** on sama hea nagu teine mudel, kust see tunnus on eemaldatud.

Lõplikku mudelisse kaasati 8 tunnust. Lühiduse mõttes tähistatakse seda mudelit tähega L. Mudel L, kus $P(Segasaarlane)$ tähendab segasaarlaseks olemise tõenäosust, esitub järgmisel kujul:

$$\text{Mudel } L\text{: } \log \frac{P(Segasaarlane)}{1 - P(Segasaarlane)} = -1.44 + 1.79 * (\textbf{Vallaslaps} = \text{ei ole teada}) +$$

$$+4.17 * (\textbf{Vallaslaps} = \text{jah}) - 0.28 * \textbf{Mitmes} + 0.87 * (\textbf{Päritolu2} = \text{Vene väed}) -$$

$$-0.05 * \textbf{Pea pikkus} - 0.04 * \textbf{Sümfüüs} - 0.02 * \textbf{Puusalaius} +$$

$$+0.02 * \textbf{Käe pikkus} + 0.04 * \textbf{Jala pikkus}$$

Mudeli L põhjal on ilmselge, et kõige rohkem mõjutavad segasaarlaseks olemise shanssi tunnused **Vallaslaps**, **Päritolu2** ja **Mitmes**. Vallaslapseks olemine ja vene vägede paiknemine vallas suurendavad segasaarlaseks olemise shanssi. Kui teised tunnused on fikseeritud siis esiklapsel on kõige suurem tõenäosus olla segasaarlane ning tõenäosus väheneb iga järgmise lapse puhul. Shanss, et peres teisena sündinud laps on segasaarlane, on $e^{0.28}$ ehk 1.32 korda väiksem, kui esimesena sündinud lapse puhul. Pakub huvi, et erinevaid tunnuste kombinatsioone katsetades jäi tunnus **Sümfüüs** alati oluliseks, millest võib järeldada, et sümfüüsi kõrgus on oluline näitaja puhtsaarlaste ja segasaarlaste võrdlemisel.

Tabelis 3.3 on toodud prognoositud tulemuste ja tunnuse **Saarlane** tõeväärtuste sagedustabel. Prognoos teostati kasutades kõiki treeningandmeid, seega tabelis olevad sagedused võivad klassifitseerimisviga alahinnata [7]. Tabeli põhjal saadi mudeli L spetsiifilisuseks, st õigesti prognoositud puhtsaarlaste osakaaluks $\frac{259}{274} = 0.95$. Tundlikkuseks, st õigesti prognoositud segasaarlaste osakaaluks saadi $\frac{61}{83} = 0.73$.

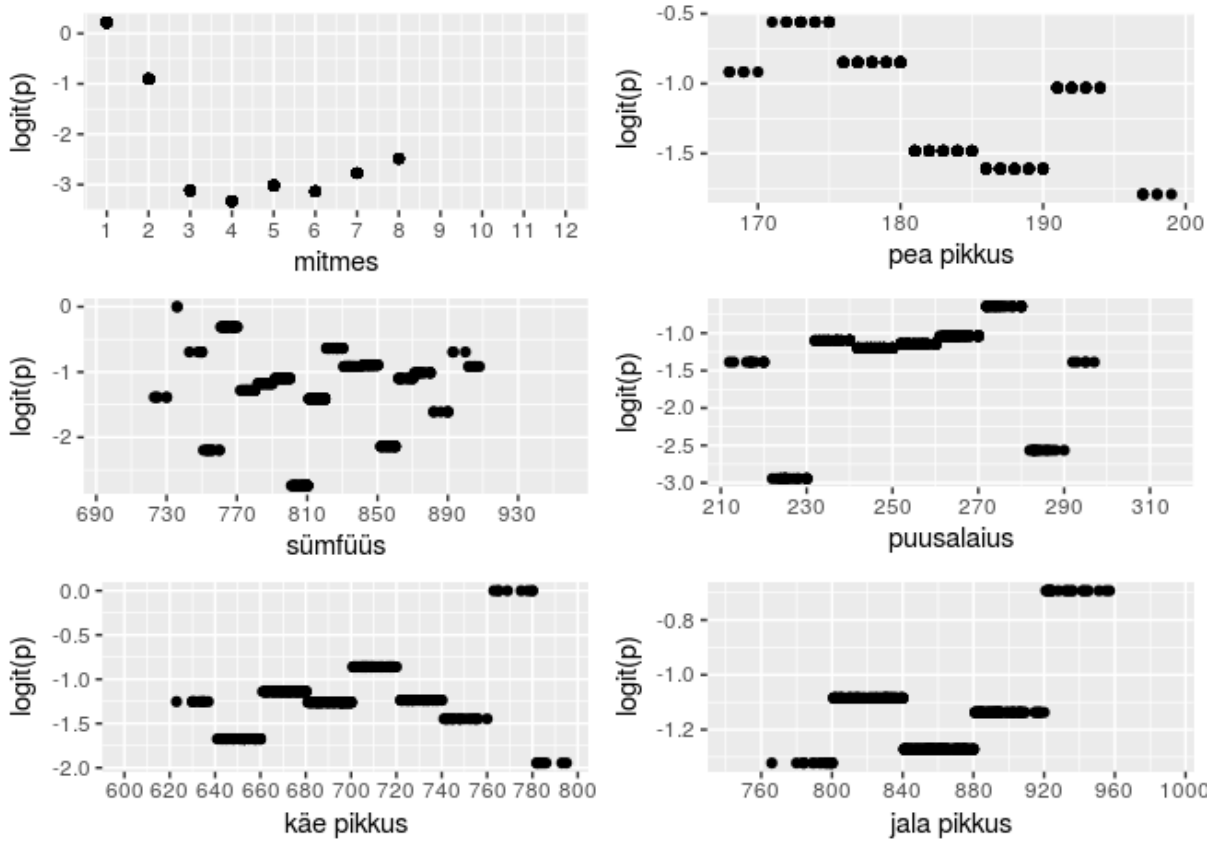
Tabel 3.3. Klassifitseerimise sagedustabel mudeli L jaoks.

Prognoos	**Tegelik olek**	
	Segasaarlane	**Saarlane**
Segasaarlane	61	15
Saarlane	22	259

Selleks, et hinnata klassifitseerimisviga kasutati K-kordset ristvalideerimist, valides K väärtuseks 5. Kuna kasutatud andmestik sisaldab vähe vaatlusi, siis suurema K väärtuse võtmine suurendaks klassifitseerimisvea hinnangu varieeruvust. Mudeli L ristvalideerimise veaks saadi 0.12. Järelikult õnnestub mudelil õigesti klassifitseerida sega- ja puhtsaarlasi hinnanguliselt 88% juhtudel.

3.4 Aditiivse logistilise mudeli rakendamine

Logistilise mudeli peamiseks eelduseks on lineaarne seos seletavate tunnuste ja huvipakkuva sündmuse toimumise tõenäosuse $logit$-funktsiooni vahel. Kui see tingimus on rikutud, siis



Joonis 3.1. Tunnuste seosed segasaarlaseks olemise log-shanssidega.

täpsema prognoosi saamiseks tuleb kasutada teisi lähenemisi, mis arvestavad andmetes esineva mittelineaarsusega. Eelnevalt vaadeldud mudel L saavutas üsna väikese klassifitseerimisvea nii treeningandmetel kui ka ristvalideerimise meetodit rakendades. Uurime, kas keerulisema mudeliga on võimalik veelgi täpsust tõsta. Esmalt uuritakse graafikute abil, kas andmetes esineb mittelineaarsus. Tunnused **Vallaslaps** ja **Päritolu2** on nominaaltunnused, mistõttu nende puhul ei ole võimalik rääkida lineaarsest seosest $logit$-funktsiooniga. Ülejäänud 6 tunnust on arvulised. Nende seos segasaarlaseks olemise hinnangulise tõenäosuse $logit$-funktsiooniga on toodud Joonisel 3.1. Kuna tunnus **Mitmes** on diskreetne tunnus, siis tõenäosuse hinnang p iga selle tunnuse fikseeritud väärtuse jaoks on segasaarlaste osakaal laste hulgas. Pidevate tunnuste puhul on tõenäosuste hindamiseks vaja jagada tunnuse väärtuspiirkond võrdseteks lõikudeks. Seejärel leitakse p iga lõigu jaoks segasaarlaste osakaaluna selles lõigus.

Joonise 3.1 põhjal näitab tunnus **Mitmes** olulist mittelineaarset seost. Esimesena sündinud lapse hinnanguline segasaarlaseks olemise log-shanss on kõige suurem. Samuti on näha, et 9.-12. sündinud laste hulgas polnud ühtegi segasaarlast. Mittelineaarset seost võib märgata ka teiste tunnuste korral. Lõplik otsus seosefunktsiooni kuju osas tehakse tarkvara abil. Aditiivse mudeli

konstrueerimisel kaasati need 8 seletavat tunnust, mida eelnevalt kasutati mudelis L. Selleks võeti kasutusele tarkvara R pakett GAM, milles on defineeritud üldistatud aditiivse mudeli funktsioon gam.

Kõigepealt alustati mudeliga, kus kõikidele arvulistele tunnustele rakendati siluvat splaini vabadusastmega 7. Tähistame seda mudelit tähega $A1$. Kõik kvalitatiivsed tunnused kaasati mudelisse $A1$ originaalkujul. Mudeli $A1$ AIC kordajaks on 202.07. Tabeli 3.4 põhjal spetsiifilisuse ja tundlikkuse väärtusteks saadi 0.96 ja 0.84. Klassifitseerimisveaks on 0.06 ehk mudel $A1$ prognoosib treeningandmetel valesti ainult 6 % juhtudel. Kõik eelnevalt nimetatud mudeli diagnostikanäitajaid arvutati, nagu mudeli L puhul, treeningandmete pealt. Ristvalideerimise tulemuseks saadi 0.15, mis näitab, et mudel $A1$ teeb vale prognoosi keskmiselt 15% juhtudel.

Tabel 3.4. Klassifitseerimise sagedustabel mudeli $A1$ jaoks.

Prognoos	**Tegelik olek**	
	Segasaarlane	**Saarlane**
Segasaarlane	70	9
Saarlane	13	265

Järgmisena vaadeldi erinevate vabadusastmetega siluvate splainide kombinatsioone ning arvestati sellega, et mõne tunnuse puhul võib lineaarne seos anda parema tulemuse. Erinevate kombinatsioonide läbivaatamisel jõuti mudelini $A2$, mille üldkuju avaldub järgmiselt:

$$\text{Mudel } A2\text{: } \log \frac{P(Segasaarlane)}{1-P(Segasaarlane)} = -2.21 + 1.66 * (\textbf{Vallaslaps} = \text{ei ole teada}) +$$

$$+4.37 * (\textbf{Vallaslaps} = jah) + s(\textbf{Mitmes}, df = 2) + 0.86 * (\textbf{Päritolu2} = \text{Vene väed}) +$$

$$+s(\textbf{Pea pikkus}, 4) + s(\textbf{Sümfüüs}, 6) + \text{s}(\textbf{Puusalaius},2) + s(\textbf{Käe pikkus}, 2) + 0.05 * \textbf{Jala pikkus}$$

On näha, et mudelis $A2$ rakendatakse lineaarset funktsiooni ainult tunnusele **Jala pikkus**. Teistele pidevatele tunnustele rakendatakse siluvaid splaine. Mudeli $A2$ AIC kordajaks saadi 193.62, mis on väiksem kui mudeli $A1$ puhul. Tabelis 3.5 on toodud prognoosi ja tegelike väärtuste sagedused. Nende põhjal saadi spetsiifilisuse ja tundlikkuse väärtuseks 0.96 ja 0.80. Ristvalideerimise klassifitseerimisviga on seekord ainult 0.09, ehk mudel $A2$ teeb keskmiselt kõige vähem vigaseid prognoose.

Tabel 3.5. Klassifitseerimise sagedustabel mudeli $A2$ jaoks.

Prognoos	**Tegelik olek**	
	Segasaarlane	**Saarlane**
Segasaarlane	66	11
Saarlane	17	263

3.5 Tulemuste võrdlus

Eelnevalt vaadeldud mudelid L, $A1$ ja $A2$ konstrueeriti kasutades kõiki käesolevaid andmeid ning samasugust tunnuste komplekti, mis võimaldab omavahel võrrelda mudelite klassifitseerimisvõimet ja sobivust andmetega. Tabelist 3.6 on näha, et treeningandmete klassifitseerimisviga on logistilise regressiooni mudelil L kõige suurem, mille põhjal võib järeldada, et võrreldes teiste mudelitega sobib mudel L andmetega kõige halvemini.

Tabel 3.6. Mudelite $A1$, $A2$ ja L headuse kriteeriumite võrdlustabel

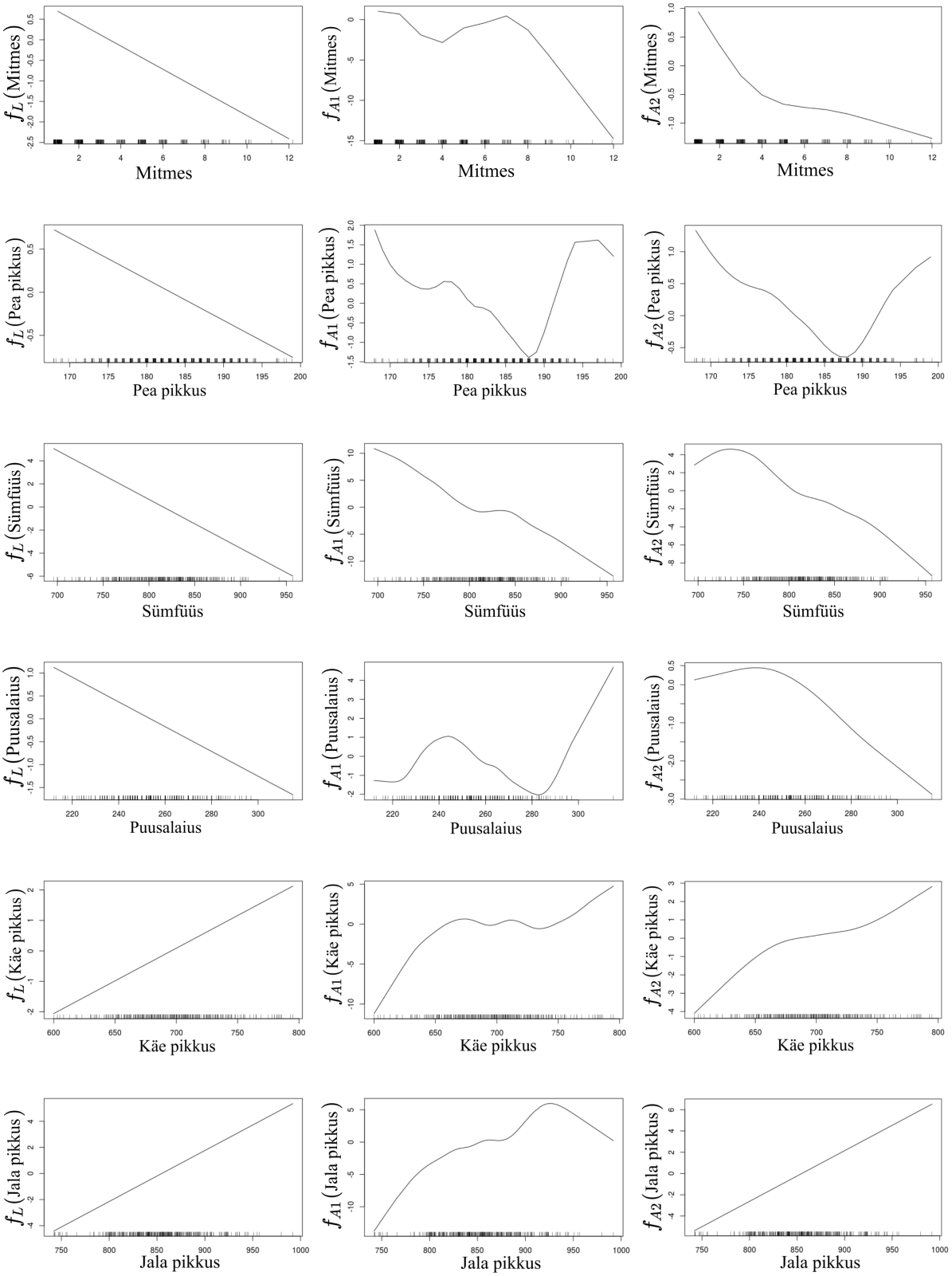
	AIC	**Spetsiifilisus**	**Tundlikus**	**Kl. viga**	**RV viga**
Mudel L	198.23	0.95	0.73	0.10	0.12
Mudel $A1$	202.07	0.96	0.84	0.06	0.15
Mudel $A2$	193.62	0.96	0.80	0.08	0.09

Klassifitseerimisviga on mudeli $A1$ jaoks 0.04 võrra väiksem kui mudeli L jaoks ehk mudel $A1$, kus igale arvulisele tunnusele rakendati siluvat splaini vabadusastmega 7, kirjeldab kasutatud andmeid paremini kui mudel L. Vaatamata madalale klassifitseerimisveale, Tabelis 3.6 on näha, et mudelil $A1$ on kõige suurem AIC väärtus ning ristvalideerimise klassifitseerimisviga on 0.15, ehk mudel $A1$ teeb keskmiselt 3% võrra rohkem vigaseid prognoose uutel andmetel, kui mudel L. Mudel $A1$ oli konstrueeritud niimoodi, et võimalikult hästi sobida andmetele, mistõttu mudel $A1$ ei üldista andmetes esinevaid seoseid ning jääb väga tundlikuks andmete käitumisele. Joonisel 3.2 on toodud mudelites L, $A1$ ja $A2$ konstrueeritud seosefunktsioonid iga seletava tunnuse ja tõenäosuse $logit$-funktsiooni vahel. On näha, et mudeliga $A1$ konstrueeritud funktsioonid on väga kõverad ning võrreldes teiste mudelitega käituvad väga sarnaselt Joonisel 3.1 toodud hinnanguliste seostega, mis põhjendab, miks mudeli $A1$ klassifitseerimisviga treeningandmetel on kõige väiksem.

Tabeli 3.6 põhjal võib teha järelduse, et mudel $A2$ on optimaalne uutel andmetel prognoosi-

miseks, kuna selle ristvalideerimise meetodil saadud prognoosivea hinnang on kõige madalam. Ristvalideerimise tulemuse kohaselt oskab mudel $A2$ õigesti klassifitseerida sega- ja puhtsaarlasi uutel andmetel keskmiselt 91% juhtudel, mis on 6% ja 3% võrra parem mudelite $A1$ ja L vastavast näitajast. Samuti on AIC kordaja mudeli $A2$ puhul kõige madalam.

Selles töös käsitleva probleemi kontekstis ei ole vahed täpsuses kriitilised, mistõttu võib eelistada logistilist mudelit selle interpreteerimise lihtsuse pärast. Ometi näitavad saadud tulemused, et arvestades nii lineaarsete kui ka mittelineaarsete seostega suudab aditiivne logistiline mudel olla suurema tulemuslikkusega kui logistiline regressioon ning selle abil võib anda täpsemat prognoosi.



Joonis 3.2. Konstrueeritud seosefunkstioonid mudeli L (vasakul), $A1$ (keskel) ja $A2$ (paremal) jaoks.

4 Kokkuvõte

Antud bakalaureusetöö eesmärgiks oli võrrelda logistilist regressioonimudelit üldistatud aditiivse mudeli erijuhuga - aditiivse logistilise regressiooniga. Võrdluse aluseks sai eeldus, et lineaarsuse nõudest loobudes annab üldistatud aditiivne mudel suurema täpsusega prognoosi. Käesoleva töö esimeses osas anti ülevaade binaarse klassifitseerimise ideest, vaadeldi kasutatud mudelite kujusid ja nende hindamise algoritme. Samuti kirjeldati kriteeriume klassifitseerimismudeli valideerimiseks, sealhulgas ka ristvalideerimise meetodit.

Mudelite võrdlus teostati Juhan Auli saarlaste antropoloogiliste mõõtmiste näitel. Antud andmed olid varasemalt Tartu Ülikooli muuseumi arhiivikogus paberkujul. Käesoleva bakalaureusetöö autor digitaliseeris need andmed. Tulemuseks on andmestik, arvutifail, mis sisaldab 57 tunnust 411 laste kohta.

Töö teises osas tutvustati andmete päritolu, kirjeldati andmete tähendust ja nendes esinevaid vigu. Lõpuks rakendati esmatöötluse läbinud andmetel logistilist regressioonimudelit ja aditiivset logistilist mudelit.

Mudelite võrdlus teostati 357 vaatlusega valimil. Analüüsi eesmärgiks püstitati klassifitseerida üldfüsioloogiliste mõõtmiste põhjal, kas laps on puhtsaarlane või segasaarlane. Segasaarlaseks peeti last, kelle ema on eestlane ja isa on venelane.

Mudeleid võrreldi mitme kriteeriume põhjal. Parim klassifitseerimise tulemus ristvalideerimise veaga 0.09 saavutati aditiivse logistilise mudeliga $A2$, kus viiele kaheksast seletavast tunnusest rakendati mittelineaarset funktsiooni.

Saadud tulemus näitas, et arvestades andmetes esineva mittelineaarsusega annab aditiivne logistiline mudel täpsema prognoosi, kui klassikaline logistiline regressioon.

Viidatud kirjandus

[1] Akaike, H., (1973). Information theory and an extension of the maximum likelihood principle, *Second international symposium on information theory*.

[2] Aul, J., (1932). Ankeedid, Saaremaa, 14-15 a, P+T. *TÜ muuseumi arhiivikogu*.

[3] Aul, J., (1933). *Maailmasõja antropoloogilisest mõjust saarlastele*, Tartu.

[4] *Eesti Entsüklopeedia*, Aul, Juhan, http://entsyklopeedia.ee/artikkel/aul_juhan, 18.03.2018.

[5] Hastie, T., Tibshirani, R., (1995). *Generalized Additive Models*, Chapman and Hall.

[6] Hastie, T., Tibshirani, R., Friedman, J., (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

[7] James, G., Witten, D., Hastie, T., Tibshirani, R., (2013). *An Introduction to Statistical Learning*, New York: springer.

[8] Kriiska, A., Tvauri, A., Selart, A., Kibal, B., Andresen, A., Pajur, A., (2013). *Eesti ajaloo atlas*, Tallinn.

[9] Käärik, E., (2017). *Andmeanalüüs II (MTMS.01.007). Loengukonspekt*.

[10] Laar, M., (2013). *20 Eesti tähtsamat lahingut*, Tallinn.

[11] Marsland, S., (2015). *Machine learning: an algorithmic perspective*, CRC press.

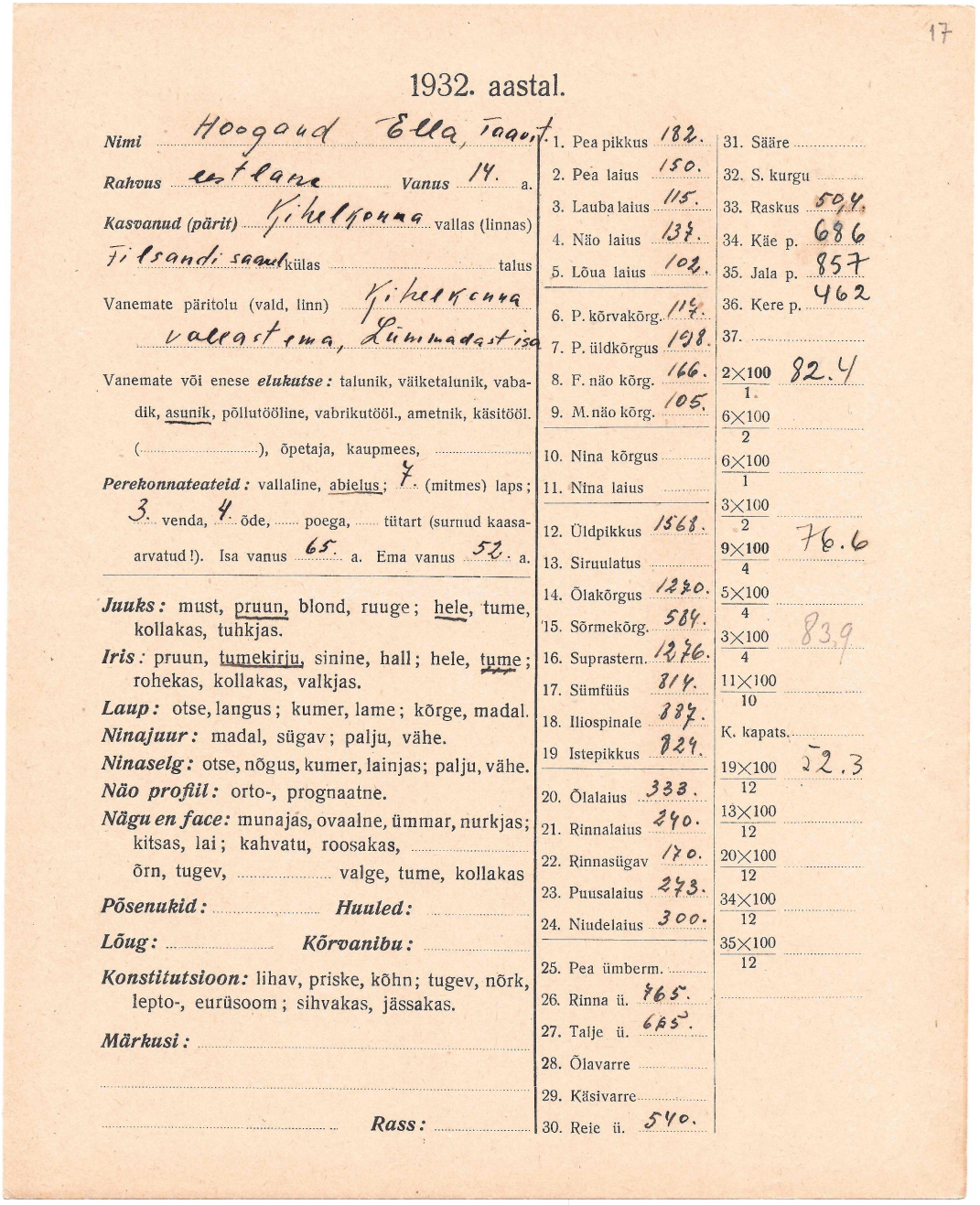
[12] Nievergelt, Y., (1993). *UMAP: Module 718; Splines in Single and Multivariable Calculus*, Lexington, MA: COMAP.

[13] Ojalo, H., (2008). *Saaremaa sõjatules. Sügis 1917*, Kuressaare.

[14] Reek, N., (1937). *Saaremaa kaitsmine ja vallutamine a. 1917*, Tallinn.

[15] Tuffry, S., (2011). *Data Mining and Statistics for Decision Making*, Chichester:Wiley.

Lisa 1. Andmeid kirjeldavad joonised



17

1932. aastal.

Nimi Hoogaud Ella, Jaagut.

Rahvus eestlane Vanus 14. a.

Kasvanud (pärit) Kihelkonna vallas (linnas)

Filsandi saare külas talus

Vanemate päritolu (vald, linn) Kihelkonna

vallast ema, Lümmadastisa

Vanemate või enese *elukutse:* talunik, väiketalunik, vabadik, asunik, põllutööline, vabrikutööl., ametnik, käsitööl. (.....), õpetaja, kaupmees, ...

Perekonnateateid: vallaline, abielus; 4. (mitmes) laps; 3. venda, 4. õde, ... poega, ... tütart (surnud kaasaarvatud!). Isa vanus 65. a. Ema vanus 52. a.

Juuks: must, pruun, blond, ruuge; hele, tume, kollakas, tuhkjas.

Iris: pruun, tumekirju, sinine, hall; hele, tume; rohekas, kollakas, valkjas.

Laup: otse, langus; kumer, lame; kõrge, madal.

Ninajuur: madal, sügav; palju, vähe.

Ninaselg: otse, nõgus, kumer, lainjas; palju, vähe.

Näo profiil: orto-, prognaatne.

Nägu en face: munajas, ovaalne, ümmar, nurkjas; kitsas, lai; kahvatu, roosakas, ... õrn, tugev, ... valge, tume, kollakas

Põsenukid: ..... ***Huuled:*** .....

Lõug: ..... ***Kõrvanibu:*** .....

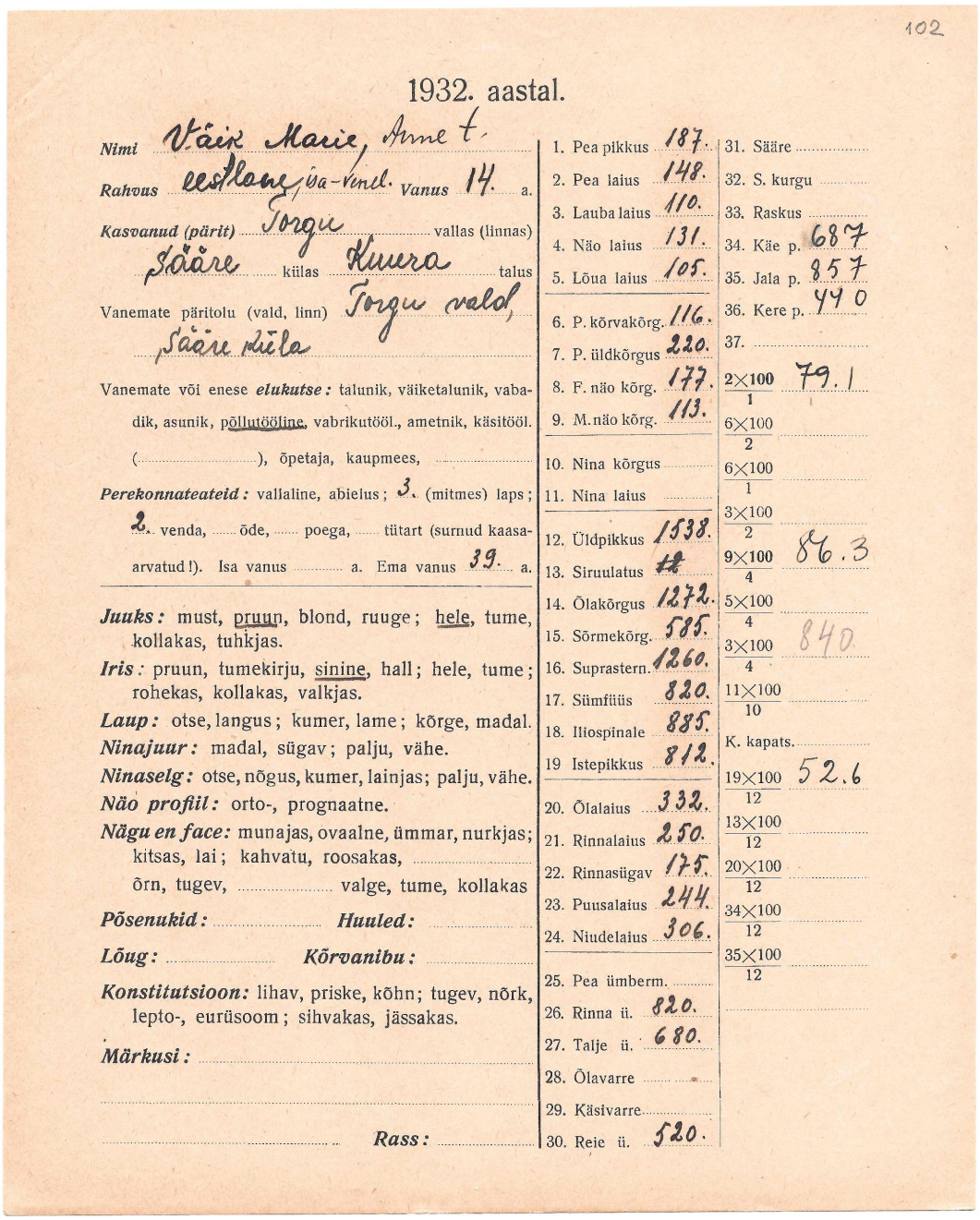
Konstitutsioon: lihav, priske, kõhn; tugev, nõrk, lepto-, eurüsoom; sihvakas, jässakas.

Märkusi: .....

Rass: .....

1. Pea pikkus	182.	31. Sääre	
2. Pea laius	150.	32. S. kurgu	
3. Lauba laius	115.	33. Raskus	50,4.
4. Näo laius	137.	34. Käe p.	686
5. Lõua laius	102.	35. Jala p.	857
6. P. kõrvakõrg.	114.	36. Kere p.	462
7. P. üldkõrgus	198.	37.	
8. F. näo kõrg.	166.	2×100/1	82.4
9. M. näo kõrg.	105.	6×100/2	
10. Nina kõrgus		6×100/1	
11. Nina laius		3×100/2	
12. Üldpikkus	1568.	9×100/4	76.6
13. Siruulatus		5×100/4	
14. Õlakõrgus	1220.	3×100/4	83,9
15. Sõrmekõrg.	584.	11×100/10	
16. Suprastern.	1276.	K. kapats.	
17. Sümfüüs	814.	19×100/12	52.3
18. Iliospinale	887.	13×100/12	
19 Istepikkus	824.	20×100/12	
20. Õlalaius	333.	34×100/12	
21. Rinnalaius	240.	35×100/12	
22. Rinnasügav	170.		
23. Puusalaius	273.		
24. Niudelaius	300.		
25. Pea ümberm.			
26. Rinna ü.	965.		
27. Talje ü.	685.		
28. Õlavarre			
29. Käsivarre			
30. Reie ü.	540.		

Joonis 4.1. J. Auli puhtsaarlase mõõtmiste ankeedi näide.



102

1932. aastal.

Nimi Väik Marie, Anne t.

Rahvus eestlane, ida-vend. Vanus 14. a.

Kasvanud (pärit) Torgu vallas (linnas)

Sääre küilas Kuura talus

Vanemate päritolu (vald, linn) Torgu vald, Sääre küla

Vanemate või enese *elukutse:* talunik, väiketalunik, vabadik, asunik, põllutööline, vabrikutööl., ametnik, käsitööl. (.....), õpetaja, kaupmees, ...

Perekonnateateid: vallaline, abielus; 3. (mitmes) laps; 2. venda, ... õde, ... poega, ... tütart (surnud kaasa-arvatud!). Isa vanus ..... a. Ema vanus 39. a.

Juuks: must, pruun, blond, ruuge; hele, tume, kollakas, tuhkjas.

Iris: pruun, tumekirju, sinine, hall; hele, tume; rohekas, kollakas, valkjas.

Laup: otse, langus; kumer, lame; kõrge, madal.

Ninajuur: madal, sügav; palju, vähe.

Ninaselg: otse, nõgus, kumer, lainjas; palju, vähe.

Näo profiil: orto-, prognaatne.

Nägu en face: munajas, ovaalne, ümmar, nurkjas; kitsas, lai; kahvatu, roosakas, ... õrn, tugev, ... valge, tume, kollakas

Põsenukid: ..... ***Huuled:*** .....

Lõug: ..... ***Kõrvanibu:*** .....

Konstitutsioon: lihav, priske, kõhn; tugev, nõrk, lepto-, eurüsoom; sihvakas, jässakas.

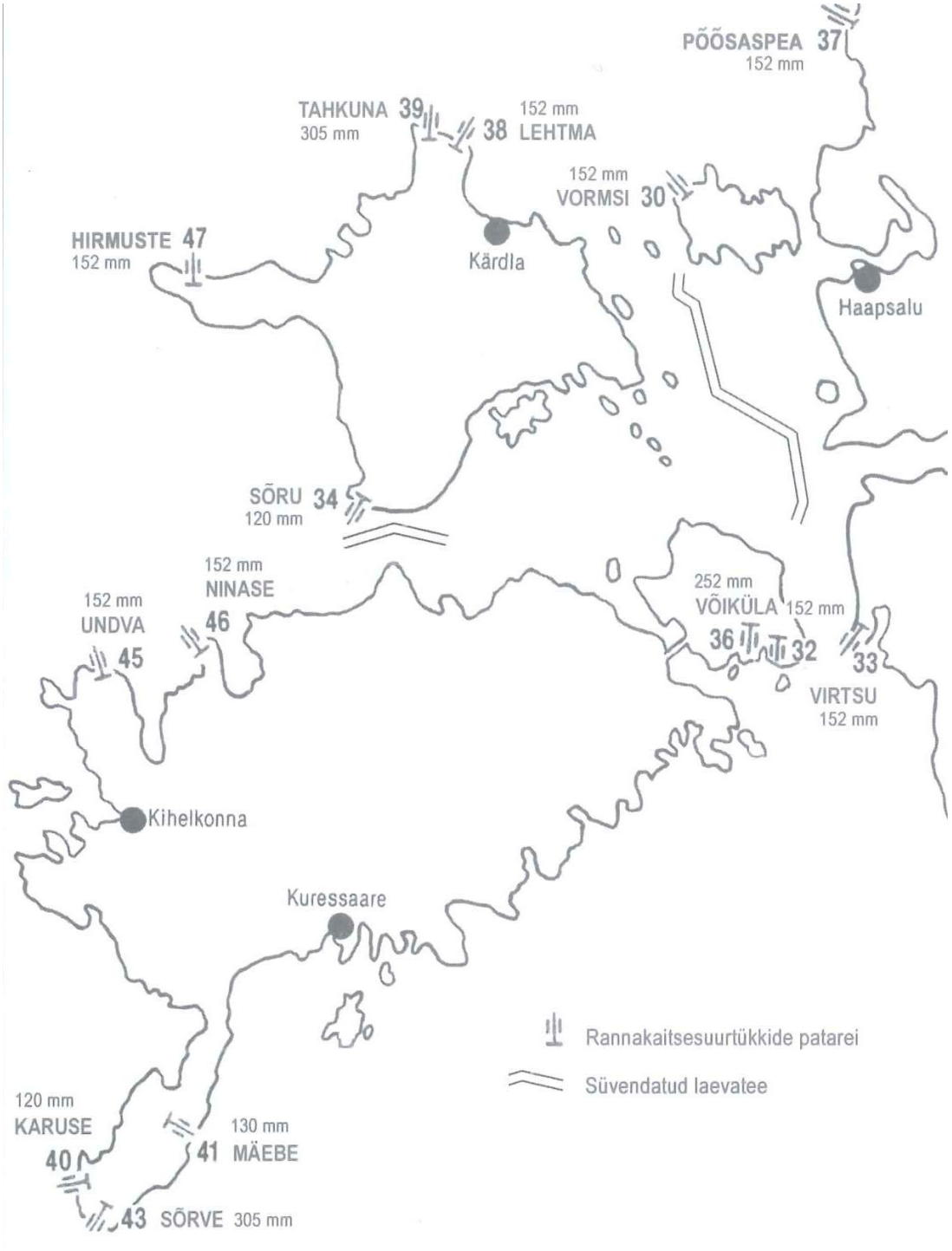
Märkusi: .....

Rass: .....

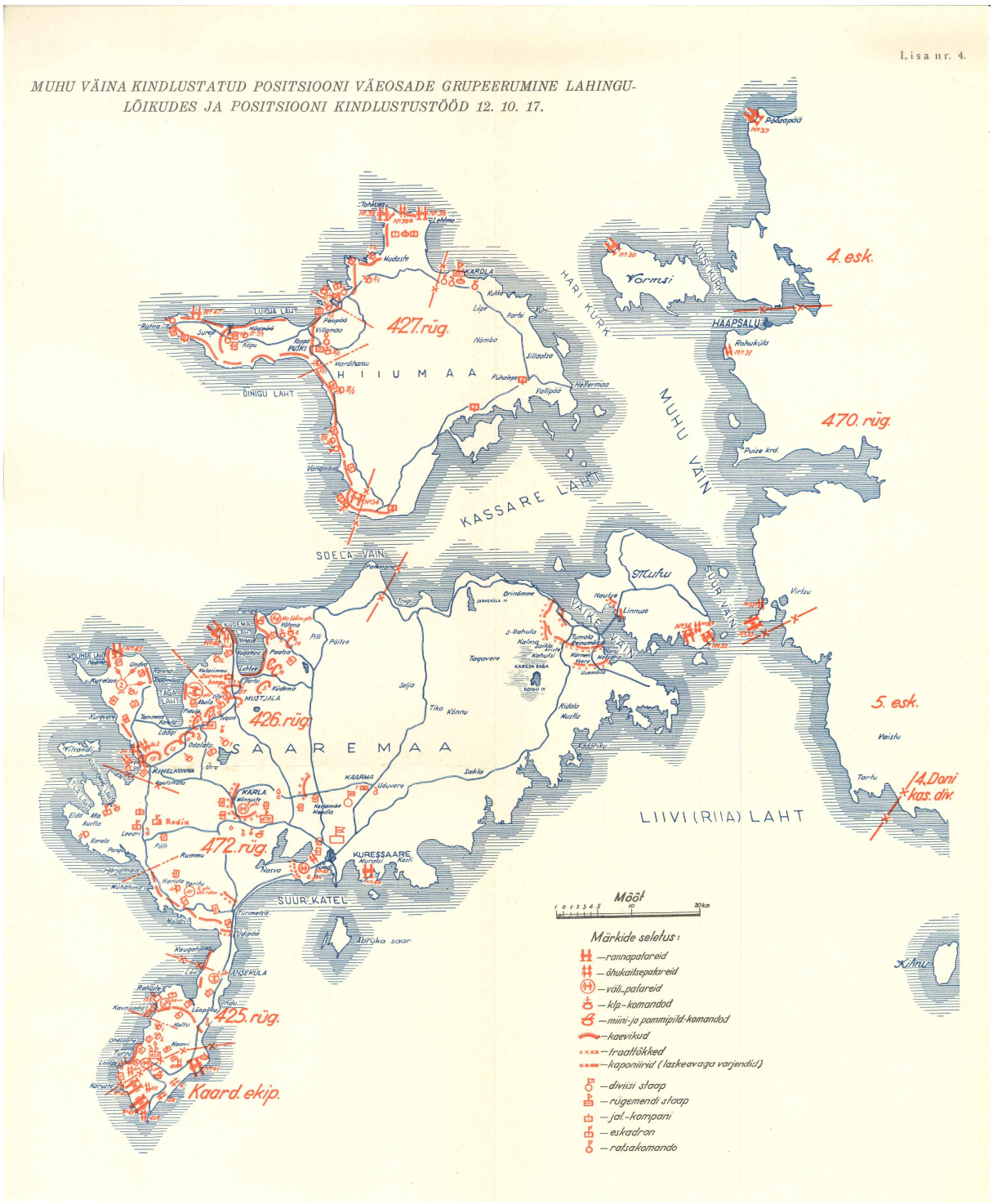
1. Pea pikkus	187.	31. Sääre	
2. Pea laius	148.	32. S. kurgu	
3. Lauba laius	110.	33. Raskus	
4. Näo laius	131.	34. Käe p.	687
5. Lõua laius	105.	35. Jala p.	857
6. P. kõrvakõrg.	116.	36. Kere p.	440
7. P. üldkõrgus	220.	37.	
8. F. näo kõrg.	177.	2×100 / 1	79.1
9. M. näo kõrg.	113.	6×100 / 2	
10. Nina kõrgus		6×100 / 1	
11. Nina laius		3×100 / 2	
12. Üldpikkus	1538.	9×100 / 4	86.3
13. Siruulatus	12	5×100 / 4	
14. Õlakõrgus	1272.	3×100 / 4	840
15. Sõrmekõrg.	585.	11×100 / 10	
16. Suprastern.	1260.	K. kapats.	
17. Sümfüüs	820.	19×100 / 12	52.6
18. Iliospinale	885.	13×100 / 12	
19 Istepikkus	812.	20×100 / 12	
20. Õlalaius	332.	34×100 / 12	
21. Rinnalaius	250.	35×100 / 12	
22. Rinnasügav	175.		
23. Puusalaius	244.		
24. Niudelaius	306.		
25. Pea ümberm.			
26. Rinna ü.	820.		
27. Talje ü.	680.		
28. Õlavarre			
29. Käsivarre			
30. Reie ü.	520.		

Joonis 4.2. J. Auli segasaarlase mõõtmiste ankeedi näide.

Joonis 4.3. Saaremaa haldusjaotus aastal 1922-1933 [8]. Kirjaga ”VeneJah” on tähistatud vallad, kus esimese maailmasõja ajal tihedalt paiknesid Vene väe kindlustustükid.



Joonis 4.4. Muhu väina kindlustatud positsiooni väeosade grupeerimine aastal 1915 [14].



Joonis 4.5. Muhu väina kindlustatud positsiooni väeosade grupeerimine aastal 1917 [13].

Lisa 2. Kasutatud R'i kood

```
#Esmane andmete töötlus
data <- data[, !names(data) \%in\%c('id', 'nimi', 'rahvus_ema', 'rahvus_isa')]

#Tunnuste juuksed Silmad ümberkodeerimine
data$juuksed<- with(data, paste0(juuksed_varv, juuksed_toon))
data$silmad<- with(data, paste0(iiris_varv,iiris_intens,iiris_toon))

data$juuksed[data$juuksed == "NANA"] <- 'NA'
data$juuksed[data$juuksed == "blondNA"] <- 'blond'
data$juuksed[data$juuksed == "NAtume"] <- 'tume'
data$juuksed[data$juuksed == "pruunNA"] <- 'pruun'
data$juuksed[data$juuksed == "ruugeNA"] <- 'ruuge'
data$juuksed[data$juuksed == "must-pruunNA"] <- 'must-pruun'
data$juuksed[data$juuksed == "must-pruun"] <- 'tume'
data$juuksed[data$juuksed == "pruuntume"] <- 'tume'
data$juuksed[data$juuksed == "pruuntumekirju"] <- 'tume'
data$juuksed[data$juuksed == "blondtume"] <- 'blond'
data$juuksed[data$juuksed == "blondhele"] <- 'blond'
data$juuksed[data$juuksed == "blondtuhkjas"] <- 'blond'
data$juuksed[data$juuksed == "pruunhele"] <- 'pruun'
data$juuksed[data$juuksed == "pruunruuge"] <- 'pruun'
data$juuksed[data$juuksed == "pruuntuhkjas"] <- 'ruuge'
data$juuksed[data$juuksed == "ruugehele"] <- 'ruuge'
data$juuksed[data$juuksed == "NA"] <- 'muu'

data$silmad[data$silmad == "sininerohekas-kollakasNA"] <- 'sininerohekas-kollakas'
data$silmad[data$silmad == "sinineNANA"] <- 'sinine'
data$silmad[data$silmad == "sinine-hallNANA"] <- 'sinine-hall'
data$silmad[data$silmad == "sinineNArohekas-kollakas"] <- 'sininerohekas-kollakas'
data$silmad[data$silmad == "pruunheleNA"] <- 'pruunhele'
data$silmad[data$silmad == "sinine-hallNArohekas"] <- 'sinine-hallrohekas'
data$silmad[data$silmad == "tumekirjutumeNA"] <- 'tumekirjutume'
data$silmad[data$silmad == "sinine-hallNArohekas"] <- 'sinine-hallrohekas'
data$silmad[data$silmad == "sinineNAkollakas"] <- 'sininekollakas'
data$silmad[data$silmad == "sinineNArohekas"] <- 'sininerohekas'
data$silmad[data$silmad == "tumekirjuNANA"] <- 'tumekirju'
data$silmad[data$silmad == "pruunNANA"] <- 'pruun'
data$silmad[data$silmad == "sinine-hallNAkollakas"] <- 'sinine-hallkollakas'
data$silmad[data$silmad == "NANANA"] <- 'NA'
data$silmad[data$silmad == "sinineNAkollakas-rohekas"] <- 'sininekollakas-rohekas'
data$silmad[data$silmad == "tumekirju-sinineNArohekas-kollakas"] <- 'tumekirju-
```

```
    sininerohekas-kollakas'
data$silmad[data$silmad == "pruunhele"] <- 'pruun'
data$silmad[data$silmad == "sininekollakas"] <- 'kirju'
data$silmad[data$silmad == "sininerohekas"] <- 'kirju'
data$silmad[data$silmad == "sininerohekas-kollakas"] <- 'kirju'
data$silmad[data$silmad == "sininetumekollakas"] <- 'kirju'
data$silmad[data$silmad == "sininekollakas-rohekas"] <- 'kirju'
data$silmad[data$silmad == "sinine-hall"] <- 'hallikas'
data$silmad[data$silmad == "sinine-hallkollakas"] <- 'hallikas'
data$silmad[data$silmad == "sinine-hallrohekas"] <- 'hallikas'
data$silmad[data$silmad == "tumekirju-sininerohekas-kollakas"] <- 'tumekirju'
data$silmad[data$silmad == "tumekirjutume"] <- 'tumekirju'
data$silmad[data$silmad == "NA"] <- 'muu'

#uus tunnus, Vene vägede asukoha põhjal
data$päritolu2 <- NA
data$päritolu2[data$päritolu == "kuressaare"] <- 'VeneJah'
data$päritolu2[data$päritolu == "kärla"] <- 'VeneJah'
data$päritolu2[data$päritolu == "abruka"] <- 'VeneJah'
data$päritolu2[data$päritolu == "kaarma"] <- 'VeneJah'
data$päritolu2[data$päritolu == "kihelkonna"] <- 'VeneJah'
data$päritolu2[data$päritolu == "lümmada"] <- 'VeneEi'
data$päritolu2[data$päritolu == "torgu"] <- 'VeneJah'
data$päritolu2[data$päritolu == "mustjala"] <- 'VeneJah'
data$päritolu2[data$päritolu == "maasi"] <- 'VeneEi'
data$päritolu2[data$päritolu == "pihtla"] <- 'VeneEi'
data$päritolu2[data$päritolu == "leisi"] <- 'VeneEi'
data$päritolu2[data$päritolu == "muhu"] <- 'VeneJah'
data$päritolu2[data$päritolu == "uuemõisa"] <- 'VeneEi'

data$mitmes[is.na(data$mitmes) & data$vallaslaps == 'jah'] <- 1
data <- data[, !names(data) %in%c('juuksed_värv', 'iiris_värv', 'juuksed_toon', '
    iiris_toon', 'iiris_intens', 'päritolu_ema', 'päritolu_isa')]
tab <- sapply(data,function(x) round(sum(is.na(x), na.rm = TRUE)/dim(data)[1],2)*
    100)
alla10 <- tab[tab < 10]
data <- data[, names(data) %in% names(alla10)]
data <- data[, !names(data) %in% c("õed", "vennad", "elukutse_v")]
data <- data[!is.na(data$mitmes),]
data$vallaslaps <- as.character(data$vallaslaps)
data$vallaslaps[is.na(data$vallaslaps)] <- 'ei ole teada'

data$lauba_laius[is.na(data$lauba_laius)] <- mean(data$lauba_laius, na.rm = TRUE)
```

```
data$näo_laius[is.na(data$näo_laius)] <- mean(data$näo_laius, na.rm = TRUE)
data$f_näokõrg[is.na(data$f_näokõrg)] <- mean(data$f_näokõrg, na.rm = TRUE)
data$lõua_laius[is.na(data$lõua_laius)] <- mean(data$lõua_laius, na.rm = TRUE)
data$üldpikkus[is.na(data$üldpikkus)] <- mean(data$üldpikkus, na.rm = TRUE)
data$õlalaius[is.na(data$õlalaius)] <- mean(data$õlalaius, na.rm = TRUE)
data$õlakõrgus[is.na(data$õlakõrgus)] <- mean(data$õlakõrgus, na.rm = TRUE)
data$sõrmekõrg[is.na(data$sõrmekõrg)] <- mean(data$sõrmekõrg, na.rm = TRUE)
data$iliospinale[is.na(data$iliospinale)] <- mean(data$iliospinale, na.rm = TRUE)
data$jala_p[is.na(data$jala_p)] <- mean(data$jala_p, na.rm = TRUE)
data$istepikkus[is.na(data$istepikkus)] <- mean(data$istepikkus, na.rm = TRUE)
data$käe_p[is.na(data$käe_p)] <- mean(data$käe_p, na.rm = TRUE)
data$rinnaümb[is.na(data$rinnaümb)] <- mean(data$rinnaümb, na.rm = TRUE)
data$suprastern[is.na(data$suprastern)] <- mean(data$suprastern, na.rm = TRUE)
data$sümfüüs[is.na(data$sümfüüs)] <- mean(data$sümfüüs, na.rm = TRUE)
data$puusalaius[is.na(data$puusalaius)] <- mean(data$puusalaius, na.rm = TRUE)
data$kere_p[is.na(data$kere_p)] <- mean(data$kere_p, na.rm = TRUE)

data$saarlane <- 1
data$saarlane[data$puhtsaarlane == 'jah'] <- 0

library(gam)
library(gtools)
library(MASS)
library(ggplot2)
library(gridExtra)

#ristvalideerimise funktsioon
CV <- function(data, formula, model_name,K = 5, tresh = 0.6) {
  indeces <- seq(1,dim(data)[1], 1)
  shnuffled <- sample(indeces)
  groups <- split(shnuffled, sample(1:K), drop = TRUE)
  summed <- 0
  for (i in 1:K){
    test <- data[unlist(groups[i]),]
    train <- data[unlist(groups[-i]),]

    if(model_name == 'log'){
      model <- glm(formula, family=binomial(link='logit'),data=train)
    }else{
      model <- gam(formula, family=binomial(link='logit'),data=train)
    }
    probs <- predict(model, test)
    pred <- rep(0, length(probs))
```

```
    pred[probs > tresh ] <- 1
    true <- test$saarlane
    summed <- summed + mean(pred == true)
  }
  return(summed/K)
}

#Logistiline mudel
log.mudel <- glm(saarlane ~sugu+vallaslaps +mitmes + vanus+päritolu2+
                   pea_pikkus+pea_laius+lauba_laius+näo_laius+lõua_laius+
                   p_kõrvakõrg+p_üldkõrg+f_näokõrg+m_näokõrg+üldpikkus+õlakõrgus+
                   sõrmekõrg+suprastern+sümfüüs+iliospinale+istepikkus+õlalaius+
                   puusalaius+rinnaümb+ käe_p+jala_p+kere_p+juuksed+silmad
                 ,family=binomial(link='logit'),data=data)
summary(log.mudel)

stepAIC <- stepAIC(log.mudel, direction = 'both')
stepAIC$anova

final_log <-glm(saarlane ~vallaslaps + mitmes + päritolu2 + pea_pikkus + p_üldkõrg +
     sümfüüs + puusalaius + käe_p + jala_p, family=binomial(link='logit'),data=data)


final_log2 <-glm(saarlane ~vallaslaps + mitmes + päritolu2 + pea_pikkus + sümfüüs +
    puusalaius + käe_p + jala_p,family=binomial(link='logit'),data=data)

final_log2$aic-final_log$aic

anova(final_log2, final_log, test="Chisq")

p <- predict(final_log2, type=c("response"))
pred_class <- rep('jah', length(p))
pred_class[p > 0.6] <- 'ei'
table(pred_class,data$puhtsaarlane)

1-CV(data, saarlane ~vallaslaps + mitmes + päritolu2 + pea_pikkus + sümfüüs +
    puusalaius + käe_p + jala_p
     ,'log', 5)

#lineaarsuse kontroll
data$count <- 1
plot_logits <- function(variable,min, max, by){
  grupid <- cut(variable, breaks=c(seq(min,max,by)))
```

```
  ss<- tapply(data$saarlane,grupid,sum)
  kokku <- tapply(data$count,grupid,sum)
  p <- ss/kokku
  l <- log(p/(1-p))
  return(l[grupid])
}

summary(data$mitmes)
out_mitmes <- plot_logits(data$mitmes, 0,12,1)
out_mitmes[out_mitmes ==  -Inf] <- NA
p1<- ggplot(data = data.frame(data$mitmes, out_mitmes), aes(x = data$mitmes, y = out
    _mitmes))+geom_point()+ scale_x_continuous(breaks = seq(1,12,1)) +labs(x = '
    mitmes',y= 'logit(p)')
p1

summary(data$pea_pikkus)
out_pea_pikkus <- plot_logits(data$pea_pikkus, 160,200,5)
out_pea_pikkus[out_pea_pikkus ==  -Inf] <- NA
p2<- ggplot(data = data.frame(data$pea_pikkus, out_pea_pikkus), aes(x = data$pea_
    pikkus, y = out_pea_pikkus))+geom_point()+scale_x_continuous(breaks = seq
    (160,200,10)) +labs(x = 'pea pikkus',y='logit(p)')
p2

summary(data$sümfüüs)
out_sümfüüs <- plot_logits(data$sümfüüs, 690,960,10)
out_sümfüüs[out_sümfüüs ==  -Inf] <- NA
p3 <- ggplot(data = data.frame(data$sümfüüs, out_sümfüüs), aes(x = data$sümfüüs, y =
     out_sümfüüs))+geom_point()+scale_x_continuous(breaks = seq(690,960,40)) +labs(x
     = 'sümfüüs',y= 'logit(p)')
p3

summary(data$puusalaius)
out_puusalaius <- plot_logits(data$puusalaius, 210,320,10)
out_puusalaius[out_puusalaius ==  -Inf] <- NA
p4<- ggplot(data = data.frame(data$puusalaius, out_puusalaius), aes(x = data$
    puusalaius, y = out_puusalaius))+geom_point()+scale_x_continuous(breaks = seq
    (210,320,20)) +labs(x = 'puusalaius',y='logit(p)')
p4

summary(data$käe_p)
out_käe_p <- plot_logits(data$käe_p, 600,800,20)
out_käe_p[out_käe_p ==  -Inf] <- NA
p5 <- ggplot(data = data.frame(data$käe_p, out_käe_p), aes(x = data$käe_p, y = out_k
```

```
    äe_p))+geom_point()+ scale_x_continuous(breaks = seq(600,800,20)) +labs(x = 'käe
     pikkus',y= 'logit(p)')
p5

summary(data$jala_p)
out_jala_p <- plot_logits(data$jala_p, 720,1000,40)
out_jala_p[out_jala_p ==  -Inf] <- NA
p6 <- ggplot(data = data.frame(data$jala_p, out_jala_p), aes(x = data$jala_p, y =
    out_jala_p))+ geom_point()+scale_x_continuous( breaks = seq(720,1000,40)) +labs(
    x = 'jala pikkus',y='logit(p)')
p6

#joonise 3.1 tekitamine
grid.arrange(p1, p2,p3,p4,p5, p6,ncol = 2)

#aditiivne mudel
gam_simple <-gam(saarlane ~  vallaslaps + s(mitmes,7) + päritolu2 + s(pea_pikkus,7)
    + s(sümfüüs,7) + s(puusalaius,7) + s(käe_p,7) + s(jala_p,7) ,family=binomial(
    link='logit'), data =data)
summary(gam_simple)
gam_simple$aic

p2 <- predict(gam_simple, type=c("response"))
pred_class2 <- rep('jah', length(p2))
pred_class2[p2 > 0.6] <- 'ei'
table(pred_class2, data$puhtsaarlane)

1-CV(data, saarlane ~  vallaslaps + s(mitmes,7) + päritolu2 + s(pea_pikkus,7) + s(sü
    mfüüs,7) +s(puusalaius,7) + s(käe_p,7) + s(jala_p,7),'gam', 5)

gam_final <-gam(saarlane ~  vallaslaps + s(mitmes, 2) + päritolu2 +s(pea_pikkus, 4)+
     s(sümfüüs, 6) + s(puusalaius,2) +s(käe_p,2) + jala_p ,family=binomial(link='
    logit'), data =data)
summary(gam_final)
gam_final$aic

p3 <- predict(gam_final, type=c("response"))
pred_class3 <- rep('jah', length(p3))
pred_class3[p3 > 0.6] <- 'ei'
table(pred_class3, data$puhtsaarlane)

1-CV(data, saarlane ~  vallaslaps + s(mitmes, 2) + päritolu2 + s(pea_pikkus, 4) + s(
    sümfüüs, 6) + s(puusalaius,2) +s(käe_p,2) + jala_p, 'gam', 5)
```

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Jelena Gorbova**,

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

   Logistilise ja aditiivse logistilise mudeli võrdlus saarlaste antropoloogiliste mõõtmiste näitel

   mille juhendaja on dots. Imbi Traat,

   1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 08.05.2018