

# **Luokittelumallin päätösrajojen optimointi moniluokkaisessa tapauksessa**

LAB-ammattikorkeakoulu

Tradenomi (AMK)

2024

Sipari Jesse

## Tiivistelmä

Tekijä(t) Sipari, Jesse	Julkaisun laji Opinnäytetyö, AMK	Valmistumisaika 2024
	Sivumäärä 32	
Työn nimi <b>Luokittelumallin päätösrajojen optimointi moniluokkaisessa tapauksessa</b>		
Tutkinto ja koulutusala Tradenomi, Tietojenkäsittely (AMK)		
Toimeksiantajan nimi, titteli ja organisaatio (jos opinnäytetyöllä on toimeksiantaja) Valmet Automation Oy		
Tiivistelmä <p>Tutkimuksessa tarkasteltiin päätösrajan optimointia viidelle binääriluokittelumallille, joista jokainen luokitteli eri objektiluokkaa (A, B, C, D ja E). Tutkimuksen tavoitteena oli selvittää, voidaanko päätösrajaa <math>P=0,5</math> vaihtamalla parantaa mallien suorituskykyä ja mitkä vaihtoehtoiset päätösrajat tuottaisivat parempia tilastollisesti merkitseviä tuloksia.</p> <p>Aineiston analysointi suoritettiin Python-ohjelmoinnin avulla laskemalla luokittelumalleille niiden hyvyyden mittaamiseen tarkoitetut tunnusluvut: Accuracy, Balanced Accuracy, F1 Score ja MCC. Tutkimuksessa huomioitiin tunnuslukujen keskiarvot ja 95%:n luottamusvälit.</p> <p>Balanced Accuracy -tunnusluvulla luokissa A, B, C ja D löydettiin vaihtoehtoisia päätösrajoja, joilla oli parempi suorituskyky kuin <math>P=0,5</math>. Luokassa E päätösraja <math>P=0,5</math> säilyi optimaalisena kaikilla tarkastelluilla tunnusluvuilla.</p>		
Asiasanat Hahmontunnistus, Koneoppiminen, Luokittelu, päätösraja, kuvantunnistus		

## Abstract

Author(s)	Type of Publication	Published
Jesse Sipari	Thesis. UAS	2024
	Number of Pages	
	32	
Title of Publication		
<b>Optimization of Decision Boundaries in a Multiclass Case</b>		
Degree, Field of Study		
Bachelor of Business Administration		
Organisation of the client (if the thesis work is commissioned by another party)		
Valmet Automation Oy		
Abstract		
<p>In the study, decision boundary optimization was examined for five binary classification models, each of which classified a different object class (A, B, C, D, E). The objective of the study was to determine whether adjusting the decision boundary from <math>P=0.5</math> can increase performance of the models and to identify viable alternative decision boundaries that would produce statistically significant improvements.</p> <p>The data-analysis was managed using Python programming by calculating performance metrics for the classification models, including Accuracy, Balanced Accuracy, F1 Score, and MCC. The study considered the average values of the metrics as well as their 95% confidence intervals.</p> <p>For the Balanced Accuracy metric, alternative decision boundaries with better performance than <math>P=0.5</math> were identified for classes A, B, C, and D. In class E, the <math>P=0.5</math> threshold remained optimal for all examined performance metrics.</p>		
Keywords		
Pattern Recognition, Machine Learning, Classification, Decision Boundary, Image Recognition		

## Sisällys

1	Tutkimuksen kuvaus.....	1
1.1	Johdanto .....	1
1.2	Aiheen valinta ja toimeksiantajan esittely .....	2
1.3	Tutkimuksen kuvaus ja tutkimusmenetelmät .....	3
2	Luokittelumallit koneoppimisessa .....	5
2.1	Koneoppiminen .....	5
2.2	Ohjattu oppiminen.....	5
2.2.1	Regressiomallit.....	6
2.2.2	Luokittelumallit .....	7
2.3	Neuroverkot .....	9
2.4	Luokittelumallin valinta, arviointi ja optimointi .....	10
2.4.1	Mallin valinta .....	10
2.4.2	Mallin arviointi .....	12
3	Tutkimus.....	15
3.1	Tutkimusprosessi .....	15
3.2	Aineiston kuvaus.....	15
3.3	Aineiston käsittely .....	16
3.4	Tilastollinen tarkastelu .....	19
3.5	Johtopäätökset.....	26
4	Yhteenveto ja pohdinta.....	30
	Lähteet.....	31

## 1 Tutkimuksen kuvaus

### 1.1 Johdanto

Päätösrajaa (engl. Decision boundary) käytetään yleisesti kaikenlaisessa luokittelussa. Päätösraja on raja-arvo, joka jakaa luokiteltavan asian kahteen eri luokkaan. Lääketieteessä on pitkään käytetty kynnyсарvoja apuna diagnostiikassa ja seulontatutkimuksessa. Lääkäri voi potilaan oireiden ja muiden mitattujen arvojen perusteella parantaa diagnoosin tarkkuutta sekä tehdä päätöksiä potilaan hoidontarpeesta ja mahdollisista lisätoimenpiteistä. (Synlab)

Psykiatriassa masennuksen määrittämisen apuna voidaan käyttää BDI-kyselyä, jossa kysymysten vastausvaihtoehdot ovat pisteytetty niin että vastauksista saadulla yhteispistemäärällä tulokset luokitellaan seuraaviin eri luokkiin: Normaali (0–12 pistettä), lievä masennus (13–18 pistettä), kohtalainen tai keskivaikea masennus (19–29 pistettä) ja vaikea masennus (30 pistettä tai yli) (Duodecim Käypä Hoito)

Päätösrajana 0,5 on yleisesti käytössä, mutta sen käyttö ei aina ole hyvin perusteltua, sillä on hyvin tapauskohtaista, halutaanko painottaa oikein luokkaan kuuluvaksi luokiteltuja vai väärin luokkaan kuulumattomaksi luokiteltuja objekteja. 0,5:n päätösrajalla kukin havaintopiste määritellään luokkaan kuuluvaksi, jos koneoppimisalgoritmin laskema todennäköisyys (P) on suurempi tai yhtä suuri kuin 50 % ( $= P \geq 0,5$ )

Liian alhainen päätösraja lisää väärin positiivisten, eli väärin luokkaan kuuluviksi luokiteltujen havaintopisteiden määrää etenkin niissä tapauksissa, joissa mukana on ominaisuuksiltaan samankaltaisia havaintopisteitä, jotka eivät kuitenkaan luokkaan kuulu.

Koneoppimisessa päätösrajalla tarkoitetaan valittua raja-arvoa, joka määrittää havainnolle luokan algoritmin määrittelemän todennäköisyysarvon(P) mukaan. Päätösraja sekä P ovat väliltä 0–1. P:n arvo perustuu algoritmilta syötettyyn opetusdataan. Opetusdatan esimerkkien on oltava oikein luokiteltuja ja otoksen täytyy olla tarpeeksi suuri, jotta saadaan mahdollisimman tarkkoja ennusteita. Päätösrajan ollessa 0,5 kaikki havainnot, jotka saavat suuremman tai yhtä suuren P:n arvon merkitään luokkaan kuuluviksi ja päätösrajan alittavat arvot luokkaan kuulumattomiksi. Koneoppimisen avulla voidaan esimerkiksi toteuttaa algoritmi, joka tunnistaa kuvassa olevia objekteja. (Nummenmaa 2021, 585–586, 595)

Kvantantunnistuksen avulla kuvista on mahdollista luokitella erilaisia hahmoja kuten esimerkiksi kasvoja. Luokittelua varten hahmoista valitaan niiden ominaispiirteet, joiden avulla ne on mahdollista luokitella eri luokkaan kuuluviksi. Piirteet valitaan tunnistettavien hahmojen mukaan ja luokitellussa käytettyjen piirteiden määrä on riippuvainen tunnistettavien

hahmojen monimutkaisuudesta sekä muista kuvien sisältämistä hahmoista, joilla voi olla hyvin paljon samanlaisia piirteitä, mutta kuuluvat luokkamäärityksen mukaan toiseen luokkaan. Esimerkiksi koirarotuja luokittelevan mallin on huomioitava erilaisia piirteitä, kun sellaisen mallin, joka luokittelee, onko jokin hahmo kissa vai koira. (Nummenmaa 2021, 581–583)

Tässä opinnäytetyössä päätösrajaa ja sen vaikutuksia käsitellään kuvantunnistuksen näkökulmasta. Kuvantunnistus on koneoppimisen osa-alue, jossa käytetään apuna neuroverkkoja. Konvoluutioneuroverkot mahdollistavat haasteellistenkin tunnistustehtävien suorittamisen. Kuvantunnistusta hyödynnetään useisiin eri tarkoituksiin ja mahdollisia sovelluskohteita on paljon. (Tuominen & Neittaanmäki 2019, 10.)

## 1.2 Aiheen valinta ja toimeksiantajan esittely

Aiheen valinta ja kiinnostus aiheiseen yleisesti opinnäytetyöntekijän opintoihin kuuluvan Valmetin Automaatio -liiketoimintalinjalle suoritettuna työharjoittelun myötä. Opinnäytetyö tehdään yhteistyössä Valmet Automation Oy:n kanssa ja tutkimuksessa käytetään yrityksen kuvantunnistusdataa.

Valmet Oyj on suomalainen suuryritys, joka valmistaa sellu-, paperi- ja energiatasollisuudelle maailmanlaajuisesti sekä teknologiaa että palveluita. Valmetilla on 220 vuotta teollista historiaa. Valmetin palveluihin kuuluvat mm. kunnossapidon ulkoistus, voimalaitosten parannukset ja varaosat, mittaukset sekä automaatoratkaisut. (Valmet)

Valmetiin kuuluu neljä liiketoimintalinjaa: Palvelut, Sellu ja energia, Paperit sekä Automaatio. Valmet Automation Oy tarjoaa automaatio ja tiedonhallinta ratkaisuja pääasiassa sellu, energia-, paperi- ja prosessiteollisuuden yrityksille, sekä kaasu-, ja meriteollisuudelle. Automaatio-linjan päätuotteita ovat analysaattorit ja mittaukset sekä laadunhallintajärjestelmät ja hajautetut ohjausjärjestelmät. Valmet Automation tarjoaa asiakkailleen myös kyberturvallisuuspalveluja. (Valmet)

Työtehtäviin kuuluivat kuva-analyysiin perustuva kuitu- ja partikkelianalyysi sekä siihen liittyvä kehitystyö ja raportointi. Kehitystyössä käytettiin ohjattua koneoppimista, jossa tekoälymenetelmää opetettiin tunnistamaan kuva-aineistosta erilaisia partikkeleita muiden seasta.

Opinnäytetyössä selvitetään, onko tapauksessa käytetty 0,5:n päätösraja optimaalinen vai voidaanko luokittelumallin ennustekykä parantaa jollain muulla arvolla.

### 1.3 Tutkimuksen kuvaus ja tutkimusmenetelmät

Opinnäytetyön tutkimuskysymys on ”Onko 0,5:n päätösraja optimaalinen luokittelumallissa vai onko jokin muu raja-arvo parempi?” Tutkimus on rajattu tarkastelemaan päätösrajan vaikutusta toimeksiantajan binääri luokittelumalleissa. Tutkimuksen aineisto koostuu viidestä luokittelumallista, joista jokainen on tehty tunnistamaan eri objektiluokkaa. Tutkimuksessa vertaillaan 0,5 päätösrajaa muihin mahdollisiin raja-arvoihin 0–1 väliltä. Tutkimus on tyypiltään kvantitatiivista ja tutkimuksessa käytetään tilastollisen testauksen menetelmiä selvittämään, onko vaihtoehtoisten raja-arvojen suorituskyvillä tilastollisesti merkitsevää eroa suhteessa 0,5 päätösrajaan.

Eri päätösrajojen vaikutusta luokittelumallien suorituskyvyn tarkastellaan käyttäen tunnuslukuja kuten: Accuracy, balanced accuracy,  $F_1$  Score sekä Matthew’s correlation coefficient. Jokainen tunnusluku antaa tietynlaisen näkökulman luokittelumallin suorituskyvystä ja niiden käyttö on tapauskohtaista. Tietyn tunnusluvun käyttöön vaikuttaa esimerkiksi aineiston sisältämien havaintojen määrä sekä luokkaan ja luokkaan kuulumattomien havaintojen välinen suhde.

Hyvyyden mittarina käytetään mallin ulkoista tarkkuutta (engl. Accuracy). Tarkastukseen otetaan myös muita tunnuslukuja kuten Balanced Accuracy, joka antaa paremman kuvan niissä tapauksissa, joissa datasetti on epätasapainoinen, eli sen sisältävistä havaintopisteistä huomattavan suuri osa on jakaantunut yhteen luokkaan.

Ulkoinen tarkkuus on valittu hyvyyden mittariksi, sillä se on ollut käytössä myös alkuperäisiä malleja arvioidessa. Ulkoinen tarkkuus kuvaa oikein luokiteltujen havaintopisteiden lukumäärää suhteessa kaikkien havaintopisteiden lukumäärään.

Datan käsittelyä ja analysointia varten tehdään kaksi erillistä python skriptiä, joista ensimmäinen lukee dataseiteistä jokaisen havaintopisteen käsin tehdyn luokittelun, sekä koneoppimisalgoritmin ennusteen(P) lukuna 0–1 väliltä. Skriptillä luodaan jokaiselle datasetille nelikenttä, joka sisältää oikeiden ja väärin luokittelujen negatiivisten sekä positiivisten lukumäärät. Nelikentän arvoilla lasketaan myös tarvittavat tunnusluvut, jotka tallennetaan taulukkoon csv-muodossa.

Jotta eri päätösrajojen vaikutusta mallin hyvyyteen voidaan tarkastella, jokaisesta mallista generoidaan 100 iteraatiota käyttäen mallin generointiin sekä validointiin eri dataa eri iteraatioissa. Kullekin iteraatiolle määritellään validointidatalla ennustuksen tarkkuutta kuvaavat tunnusluvut kuten esimerkiksi tarkkuus, herkkyys ja spesifisyys. Näistä tunnusluvuista lasketaan tilastolliset keski- ja hajontaluvut.

Tilastollisissa analyysissä nollahypoteesiksi ( $H_0$ ) on asetettu päätösraja  $P=0,5$  on oikea, eli malli ennustaa parhaiten tällä raja-arvolla. Vaihtoehtoinen hypoteesi ( $H_1$ ) on, että päätösraja  $P=0,5$  ei ole paras mahdollinen raja-arvo.

Hypoteesin testaaminen tehdään laskemalla tunnuslukuista jokaiselle päätösrajalle keskiarvo, keskihajonta, keskivirhe ja 95 %:n luottamusväli. Jokaisesta tunnusluvusta laskettua luottamusvälin alarajan keskiarvoa verrataan kaikilla päätösrajoilla 0,5:n päätösrajan keskiarvoon. Tämä vertailu tehdään jokaisella luokittelumallilla. Nollahypoteesi jää voimaan, jos  $P=0,5$ :llä lasketun tuloksen ja muulla  $P$ :n arvolla lasketun tuloksen välillä ei ole tilastollisesti merkitsevää eroa.

## 2 Luokittelumallit koneoppimisessa

### 2.1 Koneoppiminen

Koneoppimisen (engl. Machine learning) avulla pyritään tuottamaan ohjelmisto, joka oppii tunnistamaan ja luokittelemaan sille syötetyn aineiston sisältöä. Koneoppimisen hyödyntäminen vaatii sen, että aineiston sisältämien havaintojen välillä on jokin yhteys, jonka avulla on mahdollista päätellä yleinen sääntö, jota voidaan käyttää uuden aineiston käsittelyssä. (Alpaydin 2021, 40.)

Kevin P. Murphy (2012, 1) määrittelee koneoppimisen joukoksi menetelmiä, jotka pystyvät automaattisesti tunnistamaan aineistosta muotoja ja ennustaa niiden avulla uutta aineistoa tai olla avuksi päätöksenteossa.

Koneoppiminen jaetaan yleensä kolmeen luokkaan: Ohjattu oppiminen (engl. Supervised learning), ohjaamaton oppiminen (engl. Unsupervised learning) sekä vahvistusoppiminen (engl. Reinforcement learning). Ohjatussa oppimisessa opetusaineisto sisältää syötepiirteet sekä tiedon valitusta kohdepiirteestä. Ohjatun oppimisen tavoitteena on opettaa algoritmillemme funktio, joka yhdistää aineiston havainnot kuvaavat syötepiirteet valittuun kohdepiirteeseen arvoon. (Kelleher & Tierney 2021, 100.) Ohjaamattomassa oppimisessa opetusaineisto koostuu syötepiirteistä ilman kohdepiirteiden arvoja. Eräs esimerkki ohjaamattomasta oppimisesta on ryvästäminen (engl. Clustering). Ryvästämisessä tavoitteena on luoda havainnoista ryhmiä, joiden sisältämät havainnot ovat toisten samaan ryhmään kuuluvien havaintojen kanssa samankaltaisia ja eroavat muissa ryhmissä esiintyvistä havainnoista. (Bishop 2006, 3.) Vahvistusoppimisessa malli oppii vuorovaikutuksesta ympäristönsä kanssa saamalla joko positiivista tai negatiivista palautetta suorittaessaan toimintoja. Positiivisella palautteella vahvistetaan siihen johtanutta toimintatapaa. Malli pyrkii näin löytämään toimintatavan, jonka avulla voidaan parhaiten suorittaa annettu tehtävä. (Kämäräinen 2023, 152–153)

Tässä luvussa keskitytään ohjattuun koneoppimiseen sekä siihen liittyviin käsitteisiin ja metodeihin, sillä tämä koneoppimisen osa-alue on olennainen osa opinnäytetyön aihetta.

### 2.2 Ohjattu oppiminen

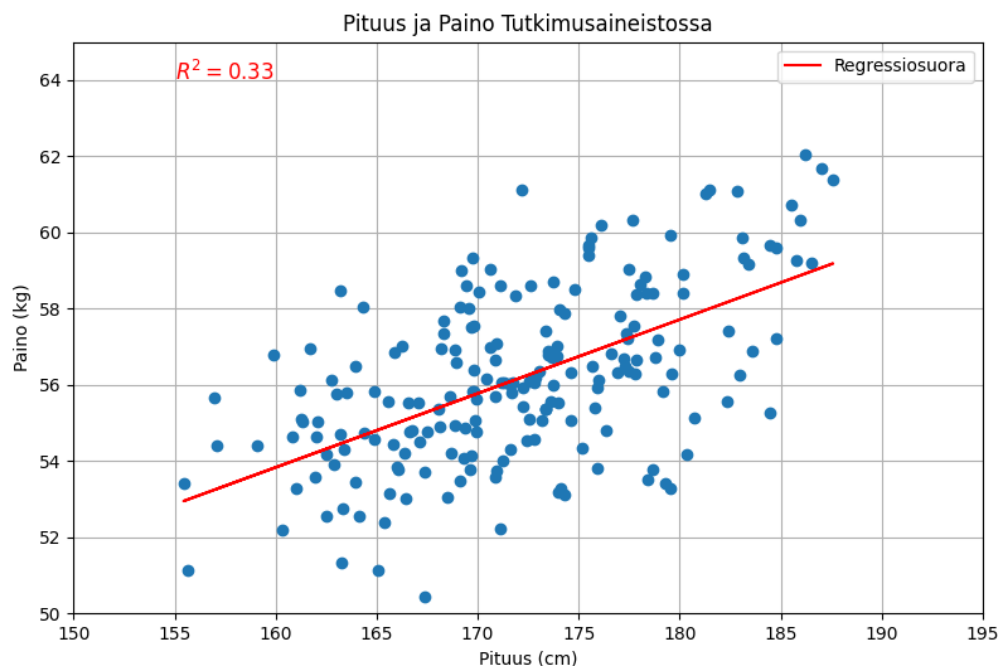
Ohjatussa oppimisessa koneoppimisalgoritmia opetetaan aineistolla, joka sisältää luokittelussa käytettävien syötepiirteet sekä ulostulostiedon kuten kohdepiirteiden luokkamerkinnot. (Goodfellow ym. 2017, 105–107) Nämä luokkamerkinnot saadaan usein tarkastelemalla aineistoa havainto kerrallaan ja merkitsemällä luokkamerkinnot käsin jokaisen havainnon kohdalla. (Bishop 2006, 1) Ohjattu oppiminen voidaan jakaa kahteen pääkategoriaan:

Luokittelu ja regressio. Regressioanalyysiä käytetään, kun selitettävä muuttuja on välimatka-asteikollinen (Hastie ym. 2017, 44) Luokittelussa jokainen syötevektori kategorisoidaan tiettyyn luokkaan kuuluvaksi. (Bishop 2006, 3)

Ohjattu oppiminen toimii niin, että koneoppimisalgoritmi käy läpi useita erilaisia funktioita löytääkseen sen, joka parhaiten kuvaa syötettä ulostuloksi. Erilaiset koneoppimisalgoritmit ovat suunniteltu etsimään tai suosimaan tietynlaisia funktioita, sillä monimutkaista aineistoa käsiteltäessä mahdollisia syötteet ulostuloksi kuvaavia kuvausten yhdistelmiä on liian suuri määrä kaikkien mahdollisten funktioiden läpikäymiseen. Tätä algoritmin ominaisuutta kutsutaan algoritmin oppimisvinoumaksi. (Kelleher & Tierney 2021, 102.)

### 2.2.1 Regressiomallit

Ohjattua oppimista ja erityisesti regressiomallien periaatetta voidaan havainnollistaa lineaarisen regressioanalyysin avulla. Lineaarinen regressioanalyysi on eräs ohjatun oppimisen osa-alue, joka kuvaa kahden tai useamman jatkuvan muuttujan välistä yhteyttä. Esimerkiksi painon ja pituuden välistä yhteyttä voidaan tutkia käyttämällä aineistoa, joka sisältää useita paino ja pituus havaintopareja. Tässä esimerkissä selitettäväksi y-muuttujaksi on valittu henkilön paino ja selitettäväksi x-muuttujaksi pituus. Y-muuttujan arvojen vaihtelua pyritään selittämään x-muuttujan arvoilla. Ennen lineaarisen regressioanalyysin suorittamista on hyvä selvittää muuttujien välisen yhteyden tyyppi. Tämä käy helposti piirtämällä sirontakuvio mallissa käytettävistä muuttujista. Kuviossa 1 on esitetty sirontakuvio, jonka jokainen piste kuvaa aineiston yksittäistä havaintoa. Sirontakuvioon on myös piirretty painon ja pituuden välistä yhteyttä kuvaava regressiosuora. Regressiosuoran kaava merkitään muodossa  $y = a + bx$ . Kaavassa y on y:n ennustettu arvo, a on vakioermi, b on regressiosuoran kulmakerroin (engl. regression coefficient) ja x on selitettävän x-muuttujan arvo. (Nummenmaa 2021, 440–446.)



Kuvio 1. Kuviossa sirontakuviio ja regressiosuora

Kuviossa on myös laskettuna korrelaatiokertoimen neliö ( $R^2=0,33$ ) jonka mukaan pituus selittää 33 % painon vaihtelusta.

### 2.2.2 Luokittelumallit

Luokittelu on ohjatun koneoppimisen osa-alue, jossa aineiston havainnot kategorisoidaan kahteen tai useampaan luokkaan. Luokittelu tehdään antamalla koneoppimisalgoritmillemme aineistoa, jossa havainnot sisältävät syötepiirteiden lisäksi tiedon niiden oikeasta luokasta. Tämän aineiston avulla se tuottaa mallin, jota voidaan käyttää uuden, tuntemattoman aineiston luokitteluun. Malli määrittää datan sisältämille objekteille luokan joko suoraan tai antaa todennäköisyysluvun (p-arvo), jonka avulla dataa analysoiva henkilö voi helposti päätellä oikean luokan. (Burkov 2019, 12.)

Luokittelun tehtävänä on määrittää mihin  $k$  kategoriaan jokin syöte kuuluu. Luokitteluongelman ratkaisemiseksi koneoppimisalgoritmin avulla pyritään yleensä tuottamaan funktio  $f: \mathbb{R}^n \rightarrow \{1, \dots, k\}$ . Kun  $y = f(x)$  malli määrittää vektorilla  $x$  kuvatun syötteen luokkaan, jonka tunnisteenä on numerokoodi  $y$ . Luokittelun avulla voidaan esimerkiksi tunnistaa kuvista erilaisia objekteja, jolloin syötteenä käytetään kuvadataa ja ulostulona saadaan kuvasta tunnistettujen objektien luokkatunnukset. (Goodfellow ym. 2017, 100.)

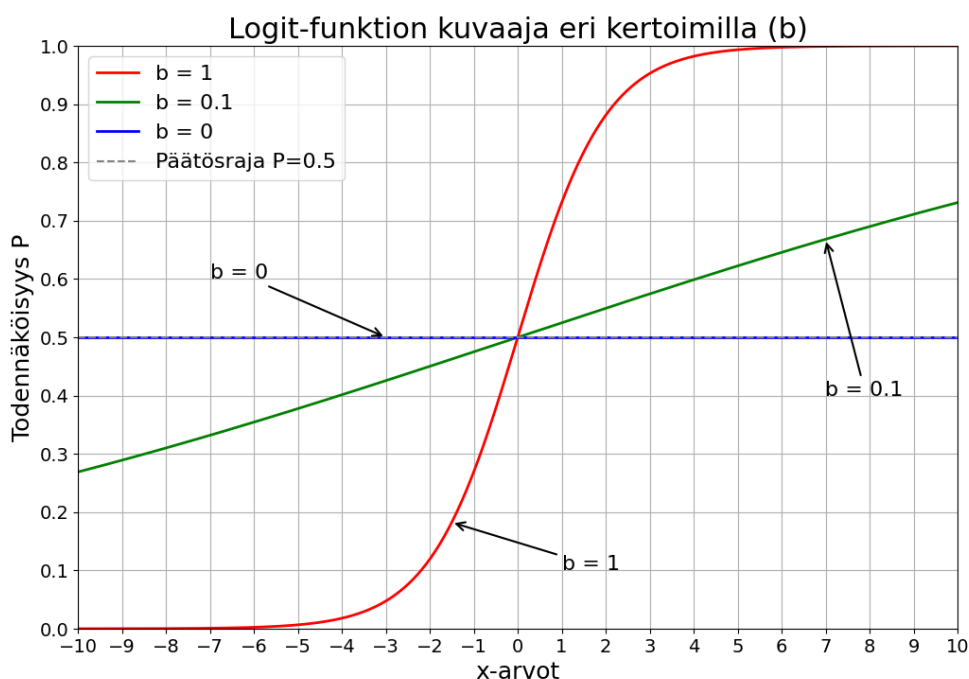
Yksi esimerkki luokittelualgoritmista on logistinen regressio. Logistinen regressio on tilastollinen menetelmä, joka soveltuu erityisesti tilanteisiin, joissa selitettävä muuttuja on dikotominen, eli sillä on kaksi mahdollista luokkaa. Esimerkiksi Nummenmaan (2021, 584) mukaan logistisen regressiomallin avulla voidaan ennustaa riskiä sairastua tyyppin 2 diabetekseen (0 = ei sairastu, 1 = sairastuu) sairauteen liittyvien riskitekijöiden lukumäärän avulla. Logistinen regressio on yleisesti käytössä lääketieteellisessä tutkimuksessa ja terveydenhuollon analyysissä esimerkin kaltaisissa tapauksissa. (Panda ym. 2022)

Logistisen regressiomallin perusideana on muodostaa matemaattinen malli, jossa selittävien muuttujien avulla ennustetaan, kuinka havainnot jakautuvat luokkiin. Logistinen regressiomalli muistuttaa lineaarista regressiomallia, jossa selitettävä muuttuja on tutkittavan tapahtuman riskin logaritmi. Mallissa muuttujien välisten yhteyksien ei oleteta olevan lineaarisia, vaan todennäköisyyksiä mallinnetaan logit -funktion kautta. Regressiomallin perusyhtälö voidaan ilmaista yhden selittävän muuttujan ja vakiotermin tapauksessa kaavalla (1). (Nummenmaa 2021, 582–586; Kaakinen & Ellonen 2023)

$$\ln\left(\frac{P}{1-P}\right) = a + bx \quad (1)$$

Yhden muuttujan tapauksessa logit -funktion kuvaaja on sigmoidi. Sigmoidi on s-kirjaimen muotoinen käyrä, jonka käyryys määräytyy selittävän muuttujan kertoimen b perusteella. Logistinen regressio etsii optimaalisen arvon kertoimelle b, tai useamman selittävän muuttujan tapauksessa, niiden kertoimille. Kuviossa 2 havainnollistetaan logistisen käyrän muotoa eri regressiokertoimilla b. Kuviossa selittävän muuttujan arvot ovat x-akselilla ja selitettävän tapahtuman todennäköisyys y-akselilla, eli tässä tapauksessa todennäköisyys millä havainto kuuluu luokkaan. Todennäköisyys eli P-arvo saa arvon nollan ja yhden väliltä. Jos kuviossa 2 valitaan esimerkkiarvot  $x = 1$  ja  $x = -5$  käyrältä, jossa  $b = 1$ , nähdään että kun  $x = 1$ , y-akselin todennäköisyys P on noin 0.73. Tämä tarkoittaa sitä, että havainnolla on noin 73 % todennäköisyys kuulua luokkaan. Arvon x ollessa -5 todennäköisyys P on noin 0.01, joten sen todennäköisyys kuulua luokkaan on vain 1 % (Kaakinen & Ellonen 2023)

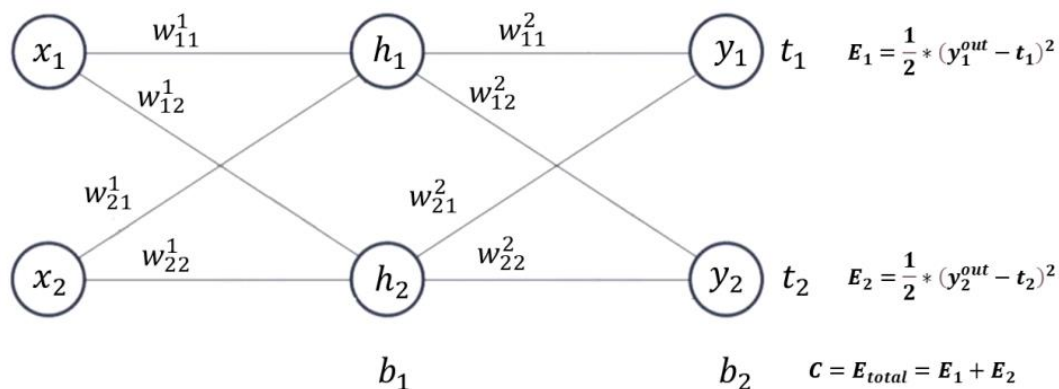
Päätös havainnon kuulumisesta positiiviseen tai negatiiviseen luokkaan eli tässä tapauksessa joko ”kuuluu luokkaan” tai ”ei kuulu luokkaan” tehdään päätösrajan avulla, joka yleensä asetetaan  $P = 0.5$ . Jos havainnon P-arvo on suurempi kuin päätösraja, niin se luokitellaan luokkaan kuuluvaksi, ja jos arvo on pienempi, havainto luokitellaan luokkaan kuulumattomaksi. Opinnäytetyön tarkoituksena on löytää tälle päätösrajalle optimaalinen arvo. Vaikka opinnäytetyössä luokittelualgoritmina on käytetty neuroverkkoa eikä logistista regressiota, niin luokan valinta tehdään samalla periaatteella eli vertaamalla havaintojen todennäköisyysarvoja päätösrajaan.



Kuvio 2. Logistinen käyrä eri regressiokertoimilla (mukailtu Kaakinen & Ellonen 2023)

### 2.3 Neuroverkot

Neuroverkkojen toiminta luokittelijana muistuttaa logistista regressiota. Kuten logistinen regressio, neuroverkko tuottaa selittävien muuttujien perusteella reaaliluvun väliltä 0–1, jota tulkitaan todennäköisyysarvona. Neuroverkoissa P-arvon laskentaprosessi tapahtuu yksinkertaisen ja vain yhden lausekkeen (logit-funktio) sijaan vaihteittain useissa toisiinsa linkittyissä yksinkertaisissa laskentasoluissa, joita kutsutaan neuroneiksi. Jokainen neuroni sisältää matemaattisen lausekkeen, joka voi olla esimerkiksi logit-funktion kaltainen.



Kuvio 3. Esimerkki neuroverkon eteenpäinlaskennasta (Uikkanen, 3)

Kuviossa 3 esitetään yksinkertainen neuroverkon rakenne, joka havainnollistaa, miten laskenta etenee tasolta toiselle. Esimerkkikuvan neuroverkko sisältää syötearvot  $x_1$  ja  $x_2$ , piilokerroksen neuronit  $h_1$  ja  $h_2$ , sekä tuloskerroksen neuronit  $y_1$  ja  $y_2$ .  $t_1$  ja  $t_2$  ovat laskennan tavoitearvot,  $w_{\text{lähtö tulo}}^{\text{kerros (iteraatio)}}$  ovat painokertoimet,  $b_1$  ja  $b_2$  ovat kerroskohtaiset vakiotermit eli bias-arvot. Uikkasen mallissa neuroniin  $h_1$  sisään tuleva arvo lasketaan kaavalla (2).

$$h_1^{\text{in}} = x_1 * w_{11}^1 + x_2 * w_{21}^1 + b_1 \quad (2)$$

Ulos lähtevä arvo lasketaan kaavalla (3).

$$h_1^{\text{out}} = \sigma(h_1^{\text{in}}) = \frac{1}{1+e^{-h_1^{\text{in}}}} \quad (3)$$

Kaavassa (3) sisään tuleva arvo sijoitetaan valittuun aktivaatiofunktioon, joka tämän esimerkin tapauksessa on logit-funktio. Laskentaa suoritetaan samalla tavalla vasemmalta oikealle, kunnes kaikille neuroneille on laskettu arvot. (Uikkanen, 2–3)

## 2.4 Luokittelumallin valinta, arviointi ja optimointi

### 2.4.1 Mallin valinta

Kun luokittelumallia toteutetaan, on päätettävä luokittelussa käytettävä algoritmi sekä etsittävä valitulle algoritmille sopivat selittävät muuttujat sekä hyperparametrien arvot, mikäli algoritmilla sellaisia on. Esimerkiksi neuroverkkoluokittelijan tapauksessa päätösraja toimii yhtenä keskeisenä hyperparametrina. Luokittelumallin ja hyperparametrien valinta tehdään vertailemalla eri vaihtoehtojen tuottamia luokittelutuloksia toisiinsa ja valitsemalla vaihtoehto, joka tarjoaa parhaan tarkkuuden luokittelulle. (Burzykowski ym. 2023, 295)

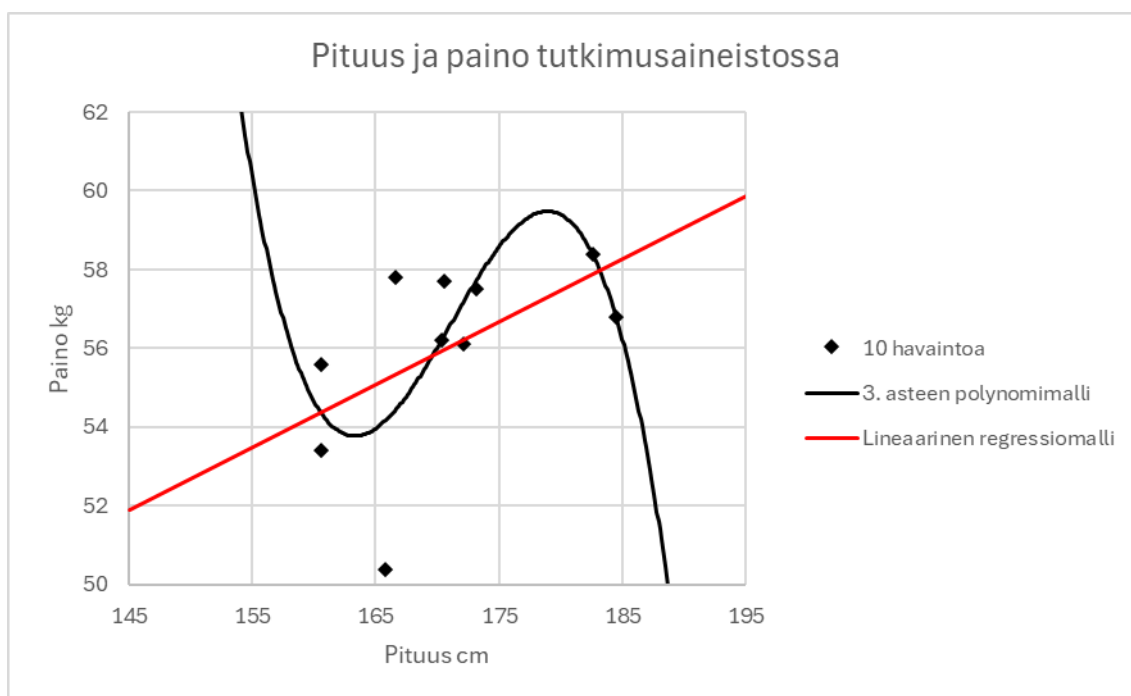
Luokittelumallin valintaan vaikuttaa sen kyky luokitella uutta aineistoa opetusaineiston pohjalta, joten eri mallien suorituskyvyn arviointi on tärkeä osa mallin valintaa tehdessä. Jos aineistoa on runsaasti, voidaan aineisto jakaa kolmeen osaan: opetusjoukko, validointijoukko ja testausjoukko. Opetusjoukkoa käytetään mallin sovittamiseen; validointijoukolla estimoidaan mallin ennustevirhettä, ja testausjoukkoa käytetään lopullisen valitun mallin yleistyskyvyn arvioimiseen. Tämä aineiston jakaminen kolmeen osaan auttaa havaitsemaan mallin mahdollisen ylisovittamisen (engl. overfitting), joka on yksi keskeisimmistä haasteista luokittelumallin toteutuksessa. (Hastie ym. 2017, 219–222)

Ylisovittaminen tarkoittaa tilannetta, jossa malli oppii seuraamaan opetusaineiston satunnaisvaihtelua sen sijaan, että oppisi aineiston yleisiä trendejä. Tämä johtaa yleensä siihen,

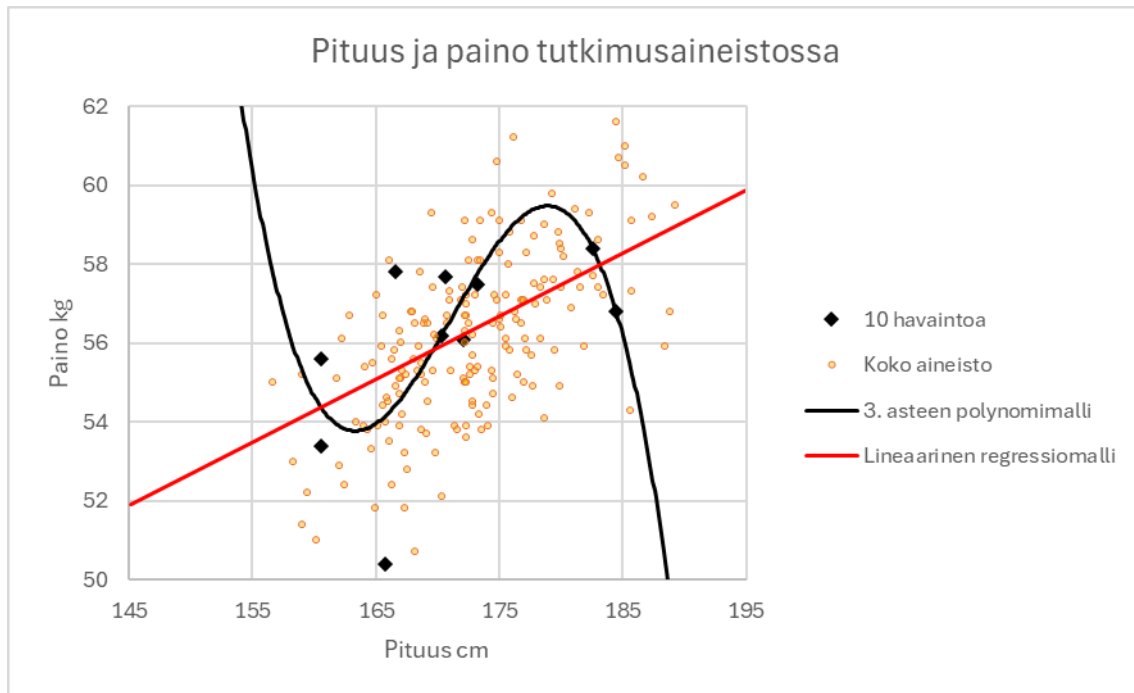
että mallin on vaikea tehdä luotettavia ennusteita uudesta aineistosta, jossa satunnaiset piirteet voivat olla erilaisia, tai puuttua kokonaan. (Burzykowski ym. 2023)

Kuviossa 4 havainnollistetaan ylisovittamista 3. asteen polynomimallin avulla. Aineistosta on valittu 10 havaintopistettä tutkimusaineistosta, joiden avulla malli ennustaa uutta aineistoa. Kuvioista 5 nähdään, että ylisovitettu malli toimii huonosti uuden aineiston kanssa.

Yksinkertaiset mallit, kuten lineaarinen regressiomalli, ovat helpommin yleistettäviä. Ne kuvaavat aineiston yleistä trendiä, mikä tekee niistä vähemmän herkkiä satunnaisvaihtelulle. Monimutkaisemmat mallit, kuten esimerkin polynomimalli, voivat näyttää tarkemmilta ope- tusaineistossa, mutta ne ovat herkempiä ylisovittamiselle, jolloin ne eivät kykene tekemään luotettavia ennusteita uudella aineistolla. Tämä korostaa yksinkertaisten mallien käyttöä erityisesti aineiston ollessa vähäinen. (Burzykowski ym. 2023)



Kuvio 4. Aineiston yleisen trendin kuvaaminen pienellä aineistolla.



Kuvio 5. Polynomimallin ja lineaarisen regressiomallin yleistettävyyden vertailu.

#### 2.4.2 Mallin arviointi

##### Luokittelutaulukko

Luokittelun suorituskykyä voidaan tarkastella luokittelutaulukon (engl. Confusion matrix) ja siitä laskettavien tunnuslukujen avulla. Luokittelutaulukko erittelee oikeat ja väärät luokitukset jokaiselle havaintopisteelle. (Zheng 2015, 8) Perinteisessä binääri luokittelutapauksessa muodostetaan  $2 \times 2$  matriisi eli nelikenttä, josta nähdään havaintopisteiden jakautuminen oikein ja väärin tunnistettuihin luokkiin, niin että niiden ennustetut luokat ovat vaakasuunnassa riveinä ja todelliset luokat pystysuunnassa sarakkeina. Joissain tapauksissa järjestys voi olla päinvastainen. 1 = luokkaan kuuluvat, 0 = ei luokkaan kuuluvat. (Kotu, 2019, 265) Taulukko 1. kuvaa luokittelutaulukkoa. (Kotu & Deshpande 2019, 264–265)

- Todelliset positiiviset (TP): Ennustettu sekä Todellinen luokka on molemmissa 1
- Väärät positiiviset (FP): Ennustettu luokka on 1 mutta todellinen luokka 0
- Väärät negatiiviset (FN): Ennustettu luokka on 0 mutta todellinen luokka 1
- Todelliset negatiiviset (TN): Ennustettu luokka sekä Todellinen luokka on molemmissa 0

Ennustettu luokka	Todellinen luokka	
	1	0
1	TP (True Positive) Oikein luokkaan kuuluviksi luokitellut objektit	FP (False Positive) Virheellisesti luokkaan kuuluviksi luokitellut objektit
0	FN (False Negative) Virheellisesti luokkaan kuulumattomiksi luokitellut objektit	TN (True Negative) Oikein luokkaan kuulumattomiksi luokitellut objektit

Mallin luokittelukyvyyn arviointiin käytetään erilaisia tunnuslukuja, jotka lasketaan luokittelutaulukon arvoilla.

Herkkyys (engl. sensitivity, recall, True positive rate) on oikeiden positiivisten ennusteiden osuus kaikista havainnoista, joiden oikea luokka on 1. Herkkyys on suhteellinen tunnusluku ja sen vaihteluväli on [0, 1]. Herkkyys lasketaan kaavalla (4). (Chicco ym. 2021)

$$(TPR) = \frac{TP}{TP+FN} \quad (4)$$

Spesifisyys (engl. specificity, true negative rate) on oikeiden negatiivisten ennusteiden osuus kaikista havainnoista, joiden oikea luokka on 0. Spesifisyys lasketaan kaavalla (5). (Chicco ym. 2021)

$$(TNR) = \frac{TN}{TN+FP} \quad (5)$$

Positiivinen ennustearvo (engl. precision, positive predictive value) on oikeiden positiivisten ennusteiden suhde kaikkiin positiivisiin ennusteisiin. (Chicco ym. 2021)

$$(PPV) = \frac{TP}{TP+FP} \quad (6)$$

Negatiivinen ennustearvo (engl. negative predictive value) on oikeiden negatiivisten ennusteiden suhde kaikkiin negatiivisiin ennusteisiin. Negatiivinen ennustearvo lasketaan kaavalla (7). (Chicco ym. 2021)

$$(NPV) = \frac{TN}{TN+FN} \quad (7)$$

Ulkoinen tarkkuus (engl. accuracy) esittää oikein luokiteltujen havaintojen suhdetta aineiston kaikkiin havaintoihin. Ulkoinen tarkkuus lasketaan kaavalla (8). (Chicco ym. 2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

Matthew's correlation coefficient lasketaan kaavalla (9). (Chicco & Jurman 2020, 4-5)

$$(MCC) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (9)$$

Balanced Accuracy on herkkyyden ja spesifisyyden aritmeettinen keskiarvo. Balanced Accuracy lasketaan kaavalla (10).

$$(BA) = \frac{TPR+TNR}{2} \quad (10)$$

F<sub>1</sub> Score on positiivisen ennustearvon ja herkkyyden harmoninen keskiarvo. (Chicco ym. 2021)

$$F_1 = 2 \frac{PPV \times TPR}{PPV + TPR} \quad (11)$$

Luokittelutaulukosta lasketuista tunnusluvuista voidaan käyttää periaatteessa mitä tahansa mallin arviointiin. Valittu tunnusluku on riippuvainen sen käyttötarkoituksesta, sillä eri tunnusluvut painottavat nelikentän eri lukuja eri tavoin.

### 3 Tutkimus

#### 3.1 Tutkimusprosessi

Tutkimuksen aineistona on viisi binääriluokittelumallia, joista jokainen on tehty tunnistamaan eri objektiluokkaa. Tässä raportissa eri luokista käytetään tunnuksia A, B, C, D ja E. Jokaisesta luokittelumallista on 100 iteraatiota, jotka on jaettu 70/30 suhteessa opetus- ja testidataksi. Datasetit sisältävät tiedot havaintojen piirteistä, Algoritmin ennustamat p-arvot sekä oikeat luokkatunnukset.

Aineiston käsittely tehtiin python koodilla, jonka avulla jokaisesta mallista ja mallin iteraatiosta laskettiin luokittelutaulukon luvut sekä tunnusluvut Accuracy, Balanced Accuracy, F1 Score, MCC, Precision, Recall sekä AUC.

Tunnusluvut laskettiin ensin päätösrajoilla väliltä 0–1 0,1:n välein ja tämän jälkeen tehtiin tarkempi tutkimus väliltä 0,2–0,8 0,01 välein. Tunnusluvuista laskettiin jokaiselle päätösrajalle keskiarvo, keskihajonta, keskivirhe ja 95 %:n luottamusväli.

Luokittelumallien suorituskyvyn analysointia jatkettiin vertaamalla kunkin tunnusluvun arvoa ja sen 95 %:n luottamusväliä eri päätösrajoilla nollahypoteesin mukaiseen päätösrajaan  $P = 0,5$ . Tunnusluvut analysoitiin tarkastelemalla, sisältyykö 0,5:n päätösrajan tunnusluvun arvo vaihtoehtoisen päätösrajan luottamusväliin. Vaihtoehtoinen päätösraja hyväksyttiin paremmaksi siinä tapauksessa, kun sen tunnusluvun arvo oli parempi kuin 0,5:n päätösrajan arvo ja luottamusväli ei sisältänyt  $P = 0,5$ :n tunnusluvun arvoa.

#### 3.2 Aineiston kuvaus

Tutkimuksen aineistona toimivat viisi eri binääriluokittelumallia, jokainen eri objektiluokkaa varten. Luokittelumallit sisältävät objektien oikean luokan, ennustetun luokan  $P$ :n arvona, sekä 22 piirrettä, joista neuroverkko on laskenut  $P$ :n arvon, joka määrittää luokan käytetyn päätösrajan mukaan. Jokainen malli on jaettu opetus- ja testidataksi suhteessa 70/30 satunnaisotantaa käyttäen. Satunnaisotanta suoritettiin 100 kertaa kullekin mallille. Taulukossa 1 on yhden luokittelumallin havaintomatriisin 20 ensimmäistä riviä ja 10 saraketta. 'Real'-sarakkeessa on havaintojen oikeat luokat ja 'est\_p'-sarakkeessa ennustettu luokka todennäköisyysarvona  $P$ . Loput sarakkeet sisältävät ne havaintojen piirteet, joista  $P$  on laskettu. Tutkimuksessa käytetään vain kahden ensimmäisen sarakkeen eli todellisen ja ennustetun luokan arvoja.

real	est_p	v1	v2	v3	v4	v5	v6	v7	v8
1	0,818	39358	39358,17	409,12	167382	122,49	15004	0,299	0,090
0	0,006	5990	77,40	110,58	12227	68,98	4758	0,624	0,389
0	0,008	4879	69,85	94,87	9001	65,48	4288	0,690	0,476
1	0,990	42197	205,42	523,30	273843	102,67	10542	0,196	0,038
1	0,932	42304	205,68	319,58	102128	168,55	28409	0,527	0,278
0	0,024	41459	203,62	286,90	82309	184,00	33856	0,641	0,411
0	0,000	3616	60,14	70,47	4966	65,34	4270	0,927	0,860
0	0,003	3839	61,96	81,36	6620	60,08	3610	0,738	0,545
1	0,989	97659	312,50	539,07	290595	230,67	53208	0,428	0,183
0	0,001	15680	125,22	328,67	108025	60,75	3690	0,185	0,034
1	0,995	119274	345,36	902,92	815258	168,20	28290	0,186	0,035
1	0,933	52361	228,83	528,75	279575	126,09	15899	0,238	0,057
0	0,296	29221	170,94	534,25	285420	69,64	4850	0,130	0,017
0	0,017	4646	68,16	104,63	10947	56,54	3197	0,540	0,292
0	0,224	38577	196,41	539,59	291160	91,03	8286	0,169	0,028
1	0,949	84781	291,17	508,24	258306	212,40	45113	0,418	0,175
1	0,819	32206	179,46	446,01	198923	91,94	8454	0,206	0,042
0	0,003	5730	75,69	100,68	10136	72,46	5251	0,720	0,518
0	0,000	14617	120,90	145,00	21026	128,35	16474	0,885	0,784
1	0,598	35236	187,71	286,74	82221	156,47	24482	0,546	0,298

Taulukko 1. Luokittelumallin havaintomatriisin 20 ensimmäistä riviä

### 3.3 Aineiston käsittely

Aineiston käsittelyä varten kehitettiin python-skripti, jonka avulla dataa voidaan analysoida. Skripti laskee luokittelumallien jokaisesta iteraatiosta luokittelutaulukon sekä tarkasteltavat tunnusluvut kaikilla päätösrajoilla 0–1 väliltä yhden desimaalin tarkkuudella ja tallentaa ne luokkakohtaisiin taulukoihin. Taulukossa 2 on erään objektiluokan 20 ensimmäistä riviä. Taulukon arvoista laskettiin päätösrajakohtaisesti jokaiselle tunnusluvulle keskiarvo, keskihajonta, vaihteluväli, keskivirhe ja 95 %:n luottamusväli erilliseen taulukkoon.

Cutoff	Iteration	TP	TN	FP	FN	Accuracy	F1 Score	BAC	MCC
0	1	377	0	525	0	0,418	0,59	0,5	0
0,1	1	376	402	123	1	0,863	0,858	0,882	0,757
0,2	1	373	444	81	4	0,906	0,898	0,918	0,824
0,3	1	370	466	59	7	0,927	0,918	0,935	0,858
0,4	1	365	482	43	12	0,939	0,93	0,943	0,878
0,5	1	362	490	35	15	0,945	0,935	0,947	0,888
0,6	1	344	504	21	33	0,94	0,927	0,936	0,877
0,7	1	317	513	12	60	0,92	0,898	0,909	0,838
0,8	1	287	518	7	90	0,892	0,855	0,874	0,787
0,9	1	227	524	1	150	0,833	0,75	0,8	0,681
1	1	0	525	0	377	0,582	0	0,5	0
0	2	376	0	526	0	0,417	0,588	0,5	0
0,1	2	374	405	121	2	0,864	0,859	0,882	0,758
0,2	2	370	436	90	6	0,894	0,885	0,906	0,802
0,3	2	363	457	69	13	0,909	0,899	0,917	0,823
0,4	2	357	470	56	19	0,917	0,905	0,922	0,834
0,5	2	350	478	48	26	0,918	0,904	0,92	0,834
0,6	2	342	490	36	34	0,922	0,907	0,921	0,841
0,7	2	329	502	24	47	0,921	0,903	0,915	0,838
0,8	2	298	510	16	78	0,896	0,864	0,881	0,789

Taulukko 2. Alkuosa taulukosta, johon on laskettu tunnusluvut eri päätösrajoilla.

Taulukossa 3 on objektiluokan E tunnuslukujen lasketut keskiarvot päätösrajoilta väliltä 0,2 – 0,8 joissa tulos on parempi kuin P=0,5 päätösrajalla. Tässä objektiluokassa kaikilla tunnusluvuilla paras arvo sijoittuu välille 0,4–0,6. Taulukossa tulosten suuruusjärjestystä on korostettu vihreällä taustavärillä. Parhaat tulokset ovat vihreällä taustalla ja vertailukohteen P=0,5 päätösrajan tulosrivit ovat lihavoitu. Taulukot luotiin jokaiselle luokalle.

Metric	Cutoff	Mean Value
<b>Accuracy</b>	<b>0,5</b>	<b>0,9268</b>
Accuracy	0,51	0,9270
Accuracy	0,52	0,9269
Accuracy	0,54	0,9271
Accuracy	0,55	0,9270
Accuracy	0,56	0,9269
Balanced Accuracy	0,41	0,9281
Balanced Accuracy	0,42	0,9281
Balanced Accuracy	0,43	0,9281
Balanced Accuracy	0,44	0,9278
Balanced Accuracy	0,45	0,9281
Balanced Accuracy	0,46	0,9282
Balanced Accuracy	0,47	0,9284
Balanced Accuracy	0,48	0,9283
Balanced Accuracy	0,49	0,9278
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,9278</b>
F1 Score	0,47	0,9143
F1 Score	0,48	0,9144
<b>F1 Score</b>	<b>0,5</b>	<b>0,9142</b>
MCC	0,48	0,8512
<b>MCC</b>	<b>0,5</b>	<b>0,8512</b>
MCC	0,51	0,8513

Taulukko 3. Luokan E tunnuslukujen parhaat keskiarvot.

Cutoff	Mean	Lower CI	Upper CI	Accepted
0,4	0,923	0,9208	0,9252	FALSE
0,41	0,9238	0,9216	0,9261	FALSE
0,42	0,9242	0,9219	0,9265	FALSE
0,43	0,9246	0,9222	0,9269	FALSE
0,44	0,9247	0,9224	0,927	FALSE
0,45	0,9252	0,9229	0,9275	FALSE
0,46	0,9258	0,9235	0,9281	FALSE
0,47	0,9263	0,924	0,9287	FALSE
0,48	0,9266	0,9243	0,9289	FALSE
0,49	0,9265	0,9242	0,9288	FALSE
<b>0,5</b>	<b>0,9268</b>	<b>0,9246</b>	<b>0,9291</b>	<b>TRUE</b>
0,51	0,927	0,9248	0,9293	FALSE
0,52	0,9269	0,9246	0,9291	FALSE
0,53	0,9268	0,9245	0,9291	FALSE
0,54	0,9271	0,9248	0,9293	FALSE
0,55	0,927	0,9247	0,9293	FALSE
0,56	0,9269	0,9246	0,9292	FALSE
0,57	0,9263	0,9239	0,9286	FALSE
0,58	0,9258	0,9235	0,9281	FALSE
0,59	0,9252	0,9229	0,9276	FALSE
0,6	0,9249	0,9227	0,9272	FALSE

Taulukko 4. Luokan E Accuracy -vertailu päätösrajoilla välillä 0,4 ja 0,6.

Taulukossa 4 esitetään luokan E testidatasta laskettujen Accuracy tunnuslukujen vertailu päätösrajoilta 0,4 ja 0,6 välillä. Ensimmäinen sarake sisältää päätösrajan, ja toisessa sarakkeessa on jokaisesta päätösrajasta laskettujen tunnuslukujen yhteenlaskettu keskiarvo. Sarakkeet "Lower CI" ja "Upper CI" kuvaavat kunkin päätösrajan 95 % luottamusväliä, jotka laskettiin tunnusluvun keskiarvon perusteella. "Lower CI" kuvaa luottamusvälin alarajaa ja "Upper CI" luottamusvälin ylärajaa. Sarake "Accepted" näyttää, täyttääkö kyseinen päätösraja kriteerit, joissa päätösrajan tunnusluvun luottamusväli ei sisällä vertailupäätösrajalle  $P=0,5$  laskettua tunnuslukua. "Accepted" saa arvon "TRUE" jos vaihtoehdoisen päätösrajan luottamusväli ei sisällä  $P=0,5$  arvoa, muuten se saa arvon "FALSE". "TRUE" tarkoittaisi sitä, että siinä tapauksessa hyväksytään vaihtoehtoinen hypoteesi.

### 3.4 Tilastollinen tarkastelu

Aineiston analysointi tehtiin vertailemalla tunnuslukujen keskiarvoja. Jokaisen tunnusluvun kohdalla verrattiin päätösrajan  $P=0,5$  keskiarvoa kaikkien muiden päätösrajojen

keskiarvoihin. Ne päätösrajat, joissa tunnusluvun keskiarvo oli suurempi tai yhtä suuri kuin  $P=0,5$  päätösrajan keskiarvo, valittiin jatkotarkasteluun (Taulukko 3).

Taulukossa 5 esitetään luokasta B lasketut tunnuslukukohtaiset keskiarvot sekä 95 %:n vaihteluväli. Taulukko sisältää vain ne päätösrajat, joilla keskiarvot ovat paremmat tai yhtä suuret kuin  $P=0,5$  päätösrajalla. Sarakkeiden selitykset ovat samat kuin taulukossa 4.

Taulukosta selviää, että luokassa B saadaan Balanced Accuracy -tunnusluvulla paremmat arvot päätösrajoilla  $P=0,4$  ja  $P=0,41$  kuin päätösrajalla  $P=0,5$ .

Luokista A, B, C ja D löytyi Balanced Accuracy -tunnusluvulle vaihtoehtoisia päätösrajoja, jotka olivat parempia kuin  $P=0,5$ . Tunnusluvuilla Accuracy, F1 Score ja MCC vaihtoehtoiset päätösrajat eivät osoittautuneet missään luokassa paremmaksi kuin  $P=0,5$  päätösraja. Luokassa E päätösraja  $P=0,5$  osoittautui parhaaksi päätösrajaksi jokaisella tunnusluvulla.

Luokkakohtaisesti tutkimustulokset olivat seuraavat:

#### Luokka A

Luokassa A (Taulukko 5) nähdään, että Balanced Accuracy reagoi tunnusluvuista eniten päätösrajan muutokseen. Tunnusluvun arvot nousevat, kun päätösrajaa lasketaan, kunnes tunnusluku saavuttaa parhaan tuloksen  $P=0,23$  päätösrajalla.  $P=0,23$  jälkeen arvot alkavat pienentyä, mutta vielä jopa  $P=0,1$  antaa paremman tuloksen kuin  $P=0,5$  päätösraja.  $P=0,5$  paremmat päätösrajat ovat väliltä  $0,1-0,41$ . Luokan A luokittelumallin tapauksessa päätösrajoja tutkittiin  $0,01-0,99$  väliltä, sillä alkuperäinen  $0,2 - 0,8$  väli ei ollut tarpeeksi löytämään kaikkia parhaita päätösrajoja.

Metric	Cutoff	Mean	Lower CI	Upper CI	Accepted
Accuracy	0,48	0,9834	0,9820	0,9848	FALSE
Accuracy	0,49	0,9834	0,9820	0,9849	FALSE
<b>Accuracy</b>	<b>0,5</b>	<b>0,9834</b>	<b>0,9819</b>	<b>0,9848</b>	<b>TRUE</b>
Accuracy	0,52	0,9834	0,9820	0,9848	FALSE
Accuracy	0,59	0,9835	0,9821	0,9849	FALSE
Accuracy	0,6	0,9834	0,9819	0,9848	FALSE
F1 Score	0,48	0,9604	0,9572	0,9637	FALSE
F1 Score	0,49	0,9604	0,9571	0,9638	FALSE
<b>F1 Score</b>	<b>0,5</b>	<b>0,9603</b>	<b>0,9569</b>	<b>0,9637</b>	<b>TRUE</b>
F1 Score	0,52	0,9603	0,9570	0,9635	FALSE
Balanced Accuracy	0,08	0,9730	0,9712	0,9748	FALSE
Balanced Accuracy	0,09	0,9737	0,9718	0,9755	FALSE
Balanced Accuracy	0,1	0,9746	0,9729	0,9764	TRUE
Balanced Accuracy	0,11	0,9752	0,9734	0,9770	TRUE
Balanced Accuracy	0,12	0,9760	0,9742	0,9778	TRUE
Balanced Accuracy	0,13	0,9764	0,9746	0,9782	TRUE
Balanced Accuracy	0,14	0,9768	0,9750	0,9786	TRUE
Balanced Accuracy	0,15	0,9771	0,9752	0,9790	TRUE
Balanced Accuracy	0,16	0,9772	0,9753	0,9791	TRUE
Balanced Accuracy	0,17	0,9772	0,9753	0,9791	TRUE
Balanced Accuracy	0,18	0,9775	0,9755	0,9794	TRUE
Balanced Accuracy	0,19	0,9775	0,9756	0,9794	TRUE
Balanced Accuracy	0,2	0,9775	0,9756	0,9794	TRUE
Balanced Accuracy	0,21	0,9775	0,9755	0,9795	TRUE
Balanced Accuracy	0,22	0,9776	0,9757	0,9796	TRUE
Balanced Accuracy	0,23	0,9779	0,9760	0,9799	TRUE
Balanced Accuracy	0,24	0,9778	0,9759	0,9798	TRUE
Balanced Accuracy	0,25	0,9778	0,9758	0,9798	TRUE
Balanced Accuracy	0,26	0,9778	0,9758	0,9798	TRUE
Balanced Accuracy	0,27	0,9777	0,9756	0,9798	TRUE
Balanced Accuracy	0,28	0,9774	0,9753	0,9796	TRUE
Balanced Accuracy	0,29	0,9775	0,9754	0,9796	TRUE
Balanced Accuracy	0,3	0,9769	0,9747	0,9792	TRUE
Balanced Accuracy	0,31	0,9769	0,9747	0,9792	TRUE
Balanced Accuracy	0,32	0,9767	0,9744	0,9790	TRUE
Balanced Accuracy	0,33	0,9764	0,9741	0,9787	TRUE
Balanced Accuracy	0,34	0,9764	0,9740	0,9787	TRUE
Balanced Accuracy	0,35	0,9760	0,9736	0,9784	TRUE
Balanced Accuracy	0,36	0,9760	0,9736	0,9783	TRUE
Balanced Accuracy	0,37	0,9759	0,9735	0,9783	TRUE
Balanced Accuracy	0,38	0,9758	0,9734	0,9782	TRUE
Balanced Accuracy	0,39	0,9756	0,9733	0,9780	TRUE
Balanced Accuracy	0,4	0,9754	0,9730	0,9777	TRUE
Balanced Accuracy	0,41	0,9753	0,9730	0,9776	TRUE
Balanced Accuracy	0,42	0,9748	0,9725	0,9772	FALSE
Balanced Accuracy	0,43	0,9748	0,9724	0,9772	FALSE
Balanced Accuracy	0,44	0,9743	0,9718	0,9768	FALSE
Balanced Accuracy	0,45	0,9741	0,9715	0,9766	FALSE
Balanced Accuracy	0,46	0,9740	0,9715	0,9766	FALSE
Balanced Accuracy	0,47	0,9739	0,9713	0,9764	FALSE
Balanced Accuracy	0,48	0,9737	0,9712	0,9762	FALSE
Balanced Accuracy	0,49	0,9732	0,9707	0,9758	FALSE
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,9728</b>	<b>0,9702</b>	<b>0,9754</b>	<b>TRUE</b>
MCC	0,48	0,9502	0,9460	0,9543	FALSE
MCC	0,49	0,9501	0,9459	0,9544	FALSE
<b>MCC</b>	<b>0,5</b>	<b>0,9500</b>	<b>0,9457</b>	<b>0,9542</b>	<b>TRUE</b>
MCC	0,52	0,9500	0,9459	0,9541	FALSE
MCC	0,59	0,9502	0,9461	0,9543	FALSE

Taulukko 5. Luokan A parhaiden päätösrajojen vertailu.

## Luokka B

Luokassa B (Taulukko 6) Balanced Accuracy on päätösrajoilla  $P=0,4$  ja  $P=0,41$  parempi kuin  $P=0,5$  päätösrajalla.  $P=0,41$  antaa Balanced Accuracyille parhaan tuloksen. Muiden tunnuslukujen kohdalla päätösraja  $P=0,5$  on optimaalinen raja-arvo.

Metric	Cutoff	Mean	Lower CI	Upper CI	Accepted
Accuracy	0,47	0,8974	0,8944	0,9004	FALSE
Accuracy	0,48	0,8976	0,8947	0,9006	FALSE
Accuracy	0,49	0,8976	0,8945	0,9006	FALSE
<b>Accuracy</b>	<b>0,5</b>	<b>0,8972</b>	<b>0,8942</b>	<b>0,9002</b>	<b>TRUE</b>
Accuracy	0,51	0,8977	0,8947	0,9007	FALSE
Accuracy	0,52	0,8979	0,8950	0,9008	FALSE
Accuracy	0,53	0,8984	0,8954	0,9013	FALSE
Accuracy	0,54	0,8984	0,8956	0,9013	FALSE
Accuracy	0,55	0,8982	0,8952	0,9011	FALSE
Accuracy	0,56	0,8975	0,8946	0,9004	FALSE
F1 Score	0,37	0,8798	0,8762	0,8833	FALSE
F1 Score	0,38	0,8807	0,8770	0,8843	FALSE
F1 Score	0,39	0,8808	0,8770	0,8846	FALSE
F1 Score	0,4	0,8812	0,8775	0,8849	FALSE
F1 Score	0,41	0,8815	0,8778	0,8852	FALSE
F1 Score	0,42	0,8812	0,8775	0,8849	FALSE
F1 Score	0,43	0,8806	0,8769	0,8844	FALSE
F1 Score	0,44	0,8803	0,8766	0,8841	FALSE
F1 Score	0,45	0,8802	0,8764	0,8839	FALSE
F1 Score	0,46	0,8802	0,8764	0,8841	FALSE
F1 Score	0,47	0,8803	0,8764	0,8841	FALSE
F1 Score	0,48	0,8802	0,8764	0,8840	FALSE
F1 Score	0,49	0,8799	0,8760	0,8838	FALSE
<b>F1 Score</b>	<b>0,5</b>	<b>0,8791</b>	<b>0,8752</b>	<b>0,8831</b>	<b>TRUE</b>
F1 Score	0,51	0,8795	0,8756	0,8834	FALSE
F1 Score	0,52	0,8794	0,8756	0,8832	FALSE
F1 Score	0,53	0,8796	0,8758	0,8834	FALSE
F1 Score	0,54	0,8794	0,8756	0,8832	FALSE
Balanced Accuracy	0,31	0,8982	0,8956	0,9008	FALSE
Balanced Accuracy	0,32	0,8983	0,8957	0,9010	FALSE
Balanced Accuracy	0,33	0,8986	0,8960	0,9012	FALSE
Balanced Accuracy	0,34	0,8991	0,8965	0,9017	FALSE
Balanced Accuracy	0,35	0,8991	0,8965	0,9017	FALSE
Balanced Accuracy	0,36	0,8995	0,8968	0,9021	FALSE
Balanced Accuracy	0,37	0,9000	0,8973	0,9026	FALSE
Balanced Accuracy	0,38	0,9007	0,8979	0,9034	FALSE
Balanced Accuracy	0,39	0,9007	0,8979	0,9035	FALSE
Balanced Accuracy	0,4	0,9010	0,898179	0,9038	TRUE
Balanced Accuracy	0,41	0,9011	0,898350	0,9039	TRUE
Balanced Accuracy	0,42	0,9008	0,8979	0,9036	FALSE
Balanced Accuracy	0,43	0,9002	0,8972	0,9031	FALSE
Balanced Accuracy	0,44	0,8998	0,8969	0,9027	FALSE
Balanced Accuracy	0,45	0,8995	0,8966	0,9025	FALSE
Balanced Accuracy	0,46	0,8995	0,8966	0,9025	FALSE
Balanced Accuracy	0,47	0,8995	0,8965	0,9025	FALSE
Balanced Accuracy	0,48	0,8993	0,8964	0,9023	FALSE
Balanced Accuracy	0,49	0,8989	0,8959	0,9020	FALSE
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,898177</b>	<b>0,8951</b>	<b>0,9012</b>	<b>TRUE</b>
Balanced Accuracy	0,51	0,8984	0,8953	0,9014	FALSE
Balanced Accuracy	0,53	0,8982	0,8952	0,9012	FALSE
MCC	0,38	0,7922	0,7865	0,7980	FALSE
MCC	0,39	0,7925	0,7866	0,7985	FALSE
MCC	0,4	0,7933	0,7874	0,7991	FALSE
MCC	0,41	0,7938	0,7880	0,7997	FALSE
MCC	0,42	0,7934	0,7875	0,7993	FALSE
MCC	0,43	0,7925	0,7865	0,7985	FALSE
MCC	0,44	0,7922	0,7862	0,7981	FALSE
MCC	0,45	0,7921	0,7861	0,7981	FALSE
MCC	0,46	0,7923	0,7863	0,7984	FALSE
MCC	0,47	0,7925	0,7864	0,7986	FALSE
MCC	0,48	0,7926	0,7866	0,7986	FALSE
MCC	0,49	0,7921	0,7860	0,7983	FALSE
<b>MCC</b>	<b>0,5</b>	<b>0,7911</b>	<b>0,7849</b>	<b>0,7973</b>	<b>TRUE</b>
MCC	0,51	0,7919	0,7857	0,7980	FALSE
MCC	0,52	0,7919	0,7859	0,7979	FALSE
MCC	0,53	0,7925	0,7865	0,7986	FALSE
MCC	0,54	0,7924	0,7864	0,7984	FALSE
MCC	0,55	0,7916	0,7855	0,7976	FALSE

Taulukko 6. Luokan B parhaiden päätösrajojen vertailu.

## Luokka C

Luokassa C (Taulukko 7) tarkasteluun otettiin, samoin kuin Luokassa A, suurempi määrä päätösrajoja. Päätösrajat valittiin väliltä 0,01 ja 0,99 sillä Balanced Accuracy antoi parempia tuloksia kuin  $P=0,5$  päätösraja kaikilla päätösrajoilla välillä 0,07–0,41. Päätösrajoilla  $P=0,2$  ja  $P=0,24$  saadaan tässä mallissa Balanced Accuracyille parhaat tulokset.

Metric	Cutoff	Mean	Lower CI	Upper CI	Accepted
Accuracy	0,47	0,9324	0,9299	0,9350	FALSE
Accuracy	0,48	0,9325	0,9300	0,9350	FALSE
Accuracy	0,49	0,9325	0,9299	0,9351	FALSE
<b>Accuracy</b>	<b>0,5</b>	<b>0,9324</b>	<b>0,9298</b>	<b>0,9350</b>	<b>TRUE</b>
Accuracy	0,51	0,9325	0,9299	0,9350	FALSE
F1 Score	0,35	0,8760	0,8710	0,8810	FALSE
F1 Score	0,38	0,8756	0,8707	0,8805	FALSE
F1 Score	0,39	0,8756	0,8707	0,8805	FALSE
F1 Score	0,4	0,8757	0,8708	0,8806	FALSE
F1 Score	0,41	0,8757	0,8709	0,8805	FALSE
F1 Score	0,44	0,8758	0,8709	0,8807	FALSE
F1 Score	0,45	0,8762	0,8714	0,8810	FALSE
F1 Score	0,46	0,8760	0,8711	0,8809	FALSE
F1 Score	0,47	0,8762	0,8713	0,8811	FALSE
F1 Score	0,48	0,8761	0,8713	0,8810	FALSE
F1 Score	0,49	0,8760	0,8710	0,8810	FALSE
<b>F1 Score</b>	<b>0,5</b>	<b>0,8756</b>	<b>0,8706</b>	<b>0,8805</b>	<b>TRUE</b>
Balanced Accuracy	0,05	0,9212	0,9176	0,9249	FALSE
Balanced Accuracy	0,06	0,9224	0,9191	0,9258	FALSE
Balanced Accuracy	0,07	0,9243	0,9211	0,9275	TRUE
Balanced Accuracy	0,08	0,9252	0,9220	0,9284	TRUE
Balanced Accuracy	0,09	0,9257	0,9225	0,9289	TRUE
Balanced Accuracy	0,1	0,9262	0,9231	0,9293	TRUE
Balanced Accuracy	0,11	0,9267	0,9236	0,9298	TRUE
Balanced Accuracy	0,12	0,9273	0,9242	0,9303	TRUE
Balanced Accuracy	0,13	0,9272	0,9241	0,9303	TRUE
Balanced Accuracy	0,14	0,9275	0,9244	0,9306	TRUE
Balanced Accuracy	0,15	0,9279	0,9248	0,9309	TRUE
Balanced Accuracy	0,16	0,9278	0,9247	0,9308	TRUE
Balanced Accuracy	0,17	0,9281	0,9251	0,9312	TRUE
Balanced Accuracy	0,18	0,9286	0,9255	0,9317	TRUE
Balanced Accuracy	0,19	0,9289	0,9258	0,9320	TRUE
Balanced Accuracy	0,2	0,9289	0,9258	0,9321	TRUE
Balanced Accuracy	0,21	0,9287	0,9255	0,9318	TRUE
Balanced Accuracy	0,22	0,9288	0,9256	0,9320	TRUE
Balanced Accuracy	0,23	0,9288	0,9256	0,9320	TRUE
Balanced Accuracy	0,24	0,9289	0,9258	0,9321	TRUE
Balanced Accuracy	0,25	0,9288	0,9256	0,9320	TRUE
Balanced Accuracy	0,26	0,9287	0,9254	0,9320	TRUE
Balanced Accuracy	0,27	0,9282	0,9248	0,9316	TRUE
Balanced Accuracy	0,28	0,9280	0,9247	0,9313	TRUE
Balanced Accuracy	0,29	0,9279	0,9246	0,9312	TRUE
Balanced Accuracy	0,3	0,9277	0,9244	0,9309	TRUE
Balanced Accuracy	0,31	0,9274	0,9241	0,9307	TRUE
Balanced Accuracy	0,32	0,9275	0,9242	0,9308	TRUE
Balanced Accuracy	0,33	0,9271	0,9238	0,9304	TRUE
Balanced Accuracy	0,34	0,9271	0,9238	0,9305	TRUE
Balanced Accuracy	0,35	0,9271	0,9238	0,9304	TRUE
Balanced Accuracy	0,36	0,9262	0,9227	0,9297	TRUE
Balanced Accuracy	0,37	0,9256	0,9221	0,9291	TRUE
Balanced Accuracy	0,38	0,9254	0,9220	0,9289	TRUE
Balanced Accuracy	0,39	0,9250	0,9216	0,9284	TRUE
Balanced Accuracy	0,4	0,9246	0,9211	0,9280	TRUE
Balanced Accuracy	0,41	0,9242	0,9207	0,9276	TRUE
Balanced Accuracy	0,42	0,9237	0,9202	0,9272	FALSE
Balanced Accuracy	0,43	0,9231	0,9195	0,9266	FALSE
Balanced Accuracy	0,44	0,9230	0,9195	0,9266	FALSE
Balanced Accuracy	0,45	0,9228	0,9192	0,9263	FALSE
Balanced Accuracy	0,46	0,9222	0,9185	0,9259	FALSE
Balanced Accuracy	0,47	0,9219	0,9183	0,9256	FALSE
Balanced Accuracy	0,48	0,9217	0,9180	0,9254	FALSE
Balanced Accuracy	0,49	0,9211	0,9174	0,9249	FALSE
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,9204</b>	<b>0,9166</b>	<b>0,9242</b>	<b>TRUE</b>
MCC	0,35	0,8305	0,8239	0,8371	FALSE
MCC	0,44	0,8302	0,8238	0,8367	FALSE
MCC	0,45	0,8309	0,8245	0,8372	FALSE
MCC	0,46	0,8306	0,8242	0,8371	FALSE
MCC	0,47	0,8310	0,8245	0,8374	FALSE
MCC	0,48	0,8309	0,8245	0,8373	FALSE
MCC	0,49	0,8307	0,8241	0,8373	FALSE
<b>MCC</b>	<b>0,5</b>	<b>0,8302</b>	<b>0,8236</b>	<b>0,8368</b>	<b>TRUE</b>

Taulukko 7. Luokan C parhaiden päätösrajojen vertailu.

## Luokka D

Myös luokan D (Taulukko 8) mallissa löytyi päätösrajoja, joissa Balanced Accuracy on  $P=0,5$  päätösrajaa parempi. Paremmat päätösrajat olivat väliltä  $P=0,33$  ja  $P=0,47$  ja parhaat keskiarvot saatiin  $P=0,41$  päätösrajalla.

Metric	Cutoff	Mean	Lower CI	Upper CI	Accepted
Accuracy	0,47	0,9377	0,9360	0,9393	FALSE
Accuracy	0,48	0,9377	0,9360	0,9394	FALSE
Accuracy	0,49	0,9377	0,9361	0,9394	FALSE
<b>Accuracy</b>	<b>0,5</b>	<b>0,9374</b>	<b>0,9357</b>	<b>0,9391</b>	<b>TRUE</b>
F1 Score	0,42	0,9129	0,9107	0,9151	FALSE
F1 Score	0,43	0,9133	0,9111	0,9155	FALSE
F1 Score	0,44	0,9132	0,9110	0,9154	FALSE
F1 Score	0,45	0,9135	0,9112	0,9157	FALSE
F1 Score	0,46	0,9136	0,9113	0,9158	FALSE
F1 Score	0,47	0,9137	0,9114	0,9160	FALSE
F1 Score	0,48	0,9135	0,9111	0,9159	FALSE
F1 Score	0,49	0,9133	0,9109	0,9156	FALSE
<b>F1 Score</b>	<b>0,5</b>	<b>0,9126</b>	<b>0,9102</b>	<b>0,9150</b>	<b>TRUE</b>
Balanced Accuracy	0,31	0,9359	0,9342	0,9376	FALSE
Balanced Accuracy	0,32	0,9365	0,9349	0,9381	FALSE
Balanced Accuracy	0,33	0,9370	0,9353	0,9386	TRUE
Balanced Accuracy	0,34	0,9374	0,9359	0,9390	TRUE
Balanced Accuracy	0,35	0,9382	0,9366	0,9397	TRUE
Balanced Accuracy	0,36	0,9386	0,9370	0,9401	TRUE
Balanced Accuracy	0,37	0,9389	0,9374	0,9405	TRUE
Balanced Accuracy	0,38	0,9391	0,9376	0,9407	TRUE
Balanced Accuracy	0,39	0,9393	0,9378	0,9409	TRUE
Balanced Accuracy	0,4	0,9393	0,9378	0,9409	TRUE
Balanced Accuracy	0,41	0,9394	0,9378	0,9410	TRUE
Balanced Accuracy	0,42	0,9393	0,9377	0,9408	TRUE
Balanced Accuracy	0,43	0,9391	0,9375	0,9407	TRUE
Balanced Accuracy	0,44	0,9385	0,9368	0,9402	TRUE
Balanced Accuracy	0,45	0,9382	0,9365	0,9399	TRUE
Balanced Accuracy	0,46	0,9378	0,9361	0,9395	TRUE
Balanced Accuracy	0,47	0,9374	0,9356	0,9392	TRUE
Balanced Accuracy	0,48	0,9369	0,9350	0,9388	FALSE
Balanced Accuracy	0,49	0,9362	0,9343	0,9381	FALSE
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,9352</b>	<b>0,9333</b>	<b>0,9371</b>	<b>TRUE</b>
MCC	0,42	0,8643	0,8610	0,8676	FALSE
MCC	0,43	0,8649	0,8616	0,8682	FALSE
MCC	0,44	0,8648	0,8614	0,8682	FALSE
MCC	0,45	0,8652	0,8618	0,8686	FALSE
MCC	0,46	0,8654	0,8620	0,8689	FALSE
MCC	0,47	0,8657	0,8622	0,8693	FALSE
MCC	0,48	0,8655	0,8618	0,8692	FALSE
MCC	0,49	0,8652	0,8616	0,8688	FALSE
<b>MCC</b>	<b>0,5</b>	<b>0,8643</b>	<b>0,8606</b>	<b>0,8679</b>	<b>TRUE</b>

Taulukko 8. Luokan D parhaiden päätösrajojen vertailu.

## Luokka E

Luokan E (Taulukko 9) tunnistaminen toimi kaikilla tunnusluvuilla parhaiten päätösrajan ollessa  $P=0,5$ . Taulukosta nähdään, että joillakin päätösrajoilla laskettujen tunnuslukujen keskiarvo oli  $P=0,5$  parempi, mutta luottamusvälejä tarkasteltaessa huomataan, että ne sisälsivät kyseiseltä tunnusluvulta lasketun  $P=0,5$  keskiarvon. Tämän takia parempia vaihtoehtisten päätösrajojen keskiarvojen paremmuutta ei pidetä tilastollisesti merkittävänä.

Metric	Cutoff	Mean	Lower CI	Upper CI	Accepted
<b>Accuracy</b>	<b>0,5</b>	<b>0,9268</b>	<b>0,9246</b>	<b>0,9291</b>	<b>TRUE</b>
Accuracy	0,51	0,9270	0,9248	0,9293	FALSE
Accuracy	0,52	0,9269	0,9246	0,9291	FALSE
Accuracy	0,54	0,9271	0,9248	0,9293	FALSE
Accuracy	0,55	0,9270	0,9247	0,9293	FALSE
Accuracy	0,56	0,9269	0,9246	0,9292	FALSE
<b>F1 Score</b>	<b>0,5</b>	<b>0,9142</b>	<b>0,9116</b>	<b>0,9168</b>	<b>TRUE</b>
F1 Score	0,47	0,9143	0,9117	0,9170	FALSE
F1 Score	0,48	0,9144	0,9117	0,9170	FALSE
<b>Balanced Accuracy</b>	<b>0,5</b>	<b>0,9278</b>	<b>0,9256</b>	<b>0,9300</b>	<b>TRUE</b>
Balanced Accuracy	0,41	0,9281	0,9260	0,9302	FALSE
Balanced Accuracy	0,42	0,9281	0,9259	0,9302	FALSE
Balanced Accuracy	0,43	0,9281	0,9259	0,9303	FALSE
Balanced Accuracy	0,44	0,9278	0,9256	0,9300	FALSE
Balanced Accuracy	0,45	0,9281	0,9259	0,9303	FALSE
Balanced Accuracy	0,46	0,9282	0,9260	0,9304	FALSE
Balanced Accuracy	0,47	0,9284	0,9262	0,9307	FALSE
Balanced Accuracy	0,48	0,9283	0,9261	0,9305	FALSE
Balanced Accuracy	0,49	0,9278	0,9256	0,9301	FALSE
<b>MCC</b>	<b>0,5</b>	<b>0,8512</b>	<b>0,8467</b>	<b>0,8556</b>	<b>TRUE</b>
MCC	0,48	0,8512	0,8466	0,8558	FALSE
MCC	0,51	0,8513	0,8467	0,8558	FALSE

Taulukko 9. Luokan E parhaiden päätösrajojen vertailu.

## 3.5 Johtopäätökset

Tutkimuksessa huomattiin, että luokkien A, B, C ja D luokittelumalleissa tunnusluku Balanced Accuracy reagoi parhaiten päätösrajan muutokseen, ja jokaisessa tapauksessa päätösrajan alentamiseen. Päätösrajaa muuttamalla ei Luokan E luokittelumallissa havaittu tilastollisesti merkittävää positiivista muutosta millään tutkimuksessa käytetyistä tunnusluvuista. Käymme vielä tarkemmin läpi jokaisen luokittelumallin tuloksia, ja niitä seikkoja, jotka olisi hyvä ottaa huomioon varsinkin, kun tarkkaillaan tunnuslukua Balanced Accuracy malleissa A, B, C ja D.

Luokan A luokittelumallissa Balanced Accuracylla saadut tulokset, on osittain selitettävissä tarkastelemalla mallin aineiston jakautumista luokkiin taulukossa 10. Taulukossa esitetään aineiston positiiviset, eli luokkaan kuuluvat (64 kappaletta), sekä negatiiviset eli luokkaan kuulumattomat objektit (266 kappaletta) ja niiden suhde prosentteina. Aineistossa luokkaan kuuluvien objektien osuus on vain 19,4 % Aineisto on siis selvästi epätasapainoinen.

Class	Count	Percentage
Positive	64	19,4
Negative	266	80,6

Taulukko 10. Luokka A: n luokittelumallin aineiston jakautuminen luokkiin.

Luokkaan A kuuluvat objektit erottuvat myös hyvin piirteiltään aineiston muista objekteista. Mikä tarkoittaa sitä, että mallin kyky luokitella on niin hyvä kaikilla päätösrajoilla, että Accuracy tunnusluvulla malli antaa päätösrajalla  $P=0,01$  tuloksen 0,8622 ja päätösrajalla  $P=0,99$  tuloksen 0,939. Balanced Accuracy antaa myös samoilla päätösrajoilla arvot 0,9114 ( $P=0,01$ ) ja 0,8566 ( $P=0,99$ ).

Luokan B mallissa aineiston jako on selvästi tasapainoisempi kuin luokan A. Aineiston havainnoista 45,9 % on luokkaan kuuluvia ja loput 54,1 % luokkaan kuuluvia. Tässä mallissa Balanced Accuracy tunnusluvulla saatiin parhaat arvot  $P=0,4$  ja  $P=0,41$  päätösrajoilla.

Cutoff	True Positive	True Negative	False Positive	False Negative
0,4	121,75	161,54	23,06	9,65
0,41	121,51	161,93	22,67	9,89
0,5	118,66	164,86	19,74	12,74

Taulukko 11. Luokka B: n luokittelutaulukon tulosten vertailu.

Taulukon 11 avulla voidaan verrata luokittelumallin päätösrajoilla  $P=0,40$ ,  $P=0,41$  ja  $P=0,5$ . Taulukko sisältää näiden päätösrajojen keskiarvot oikein ja väärin luokitelluista objekteista. Tuloksista nähdään, että päätösrajalla  $P=0,5$  malli luokittelee sekä oikeat positiiviset (True Positive) ja väärät negatiiviset hieman huonommin, mutta oikeat negatiiviset (True Negative) ja väärät positiiviset (False Positive) luokitellaan hieman paremmin kuin taulukon vaihtoehtoisissa päätösrajoissa.

Luokka C

Luokan C luokittelumallin aineistossa on myös epätasapainoa kuten Taulukko 11:sta huomataan. Tällä epätasapainolla voi olla vaikutusta Balanced Accuracy tuloksiin. Taulukossa 12 verrataan mallin opetusaineiston ja testiaineiston välistä eroa tunnuslukujen välillä

päätösräjällä  $P=0,5$ . Tässä luokassa tulokset erosivat aineistojen kesken toisistaan huomattavasti muita luokkia enemmän.

Class	Count	Percentage
Positive	82	30,3
Negative	189	69,7

Taulukko 11. Luokka C: n luokittelumallin aineiston jakautuminen luokkiin.

Metric	Teach	Test	Difference
Accuracy	0,9878	0,9324	0,0554
F1 Score	0,9773	0,8756	0,1018
Balanced Accuracy	0,9865	0,9204	0,0661
MCC	0,9691	0,8302	0,1389

Taulukko 12. Luokan C tunnuslukujen keskiarvon vertailu opetus- ja testidatan välillä  $P=0,5$ .

#### Luokka D

Taulukossa 13 nähdään luokan D luokittelumallin aineiston jakautuminen. Luokkaan kuuluvia objekteja aineistossa on 35,8 % ja luokkaan kuulumattomia 64,2 % Tässäkin mallissa Balanced Accuracy saattaa herkästi reagoida epätasapainoon aineiston jakautumisen takia.

Class	Count	Percentage
Positive	310	35,8
Negative	556	64,2

Taulukko 13. Luokka D: n luokittelumallin aineiston jakautuminen luokkiin.

#### Luokka E

Luokan E luokittelumallissa nollahypoteesi ( $P=0,5$  on paras päätösraja) piti paikkansa kaikilla tutkituilla tunnusluvuilla. Aineiston jakautuminen luokkaan kuuluviin ja luokkaan kuulumattomiin objekteihin oli myös melko tasaista kuten taulukosta 14 nähdään. Aineistossa luokkaan kuulumattomia oli 16 % enemmän kuin luokkaan kuuluvia.

Class	Count	Percentage
Positive	377	41,8
Negative	525	58,2

Tuloksista huomataan, että vain Balanced Accuracy -tunnusluku reagoi päätösräjän muutokseen positiivisesti niin, että muutos oli tilastollisesti merkitsevä. Muilla tunnusluvuilla (Accuracy, F1 Score ja MCC) päätösraja  $P=0,5$  osoittautui optimaaliseksi valinnaksi. Päätösrajojen vaihtamista on siis harkittava tapauskohtaisesti. Lisäksi huomattiin, että

epätasapainoiset aineistot saattoivat vaikuttaa Balanced Accuracyn arvoihin. Luokkien välinen epätasapaino aineistossa saattaa korostaa aliedustettujen luokkien suorituskkyä Balanced Accuracy tunnusluvulla.

#### 4 Yhteenveto ja pohdinta

Opinnäytetyön tavoitteena oli tutkia päätösrajan optimointia viidelle objektiluokalle tehdyissä luokittelumalleissa, ja selvittää, miten eri päätösrajat vaikuttavat mallien suorituskyykyyn. Tilastollisessa analyysissä nollahypoteesiksi ( $H_0$ ) asetettiin  $P=0,5$  on oikea, eli malli ennustaa oikean luokan parhaiten tällä raja-arvolla. Vaihtoehtoinen hypoteesi ( $H_1$ ) oli, että päätösraja  $P=0,5$  ei ole paras mahdollinen raja-arvo.

Hypoteesin testaaminen tehtiin vertaamalla  $P=0,5$  päätösrajan suorituskyykyä muihin päätösrajoihin tilastollisten tunnuslukujen perusteella. Jokaiselle objektiluokalle laskettiin tunnuslukujen keskiarvot, keskivirheet ja 95 %:n luottamusvälit.

Tutkimuksessa huomattiin, että luokkien A, B, C ja D luokittelumalleissa Balanced Accuracy tunnusluvulla löydettiin päätösrajoja, joilla suorituskyyky oli parempaa kuin päätösrajalla  $P=0,5$ . Luokassa E päätösraja  $P=0,5$  todettiin parhaaksi valinnaksi kaikilla tutkituilla tunnusluvuilla.

Tilastollisessa tutkimuksessa tulosten luotettavuutta paransi se, että aineisto oli valmiiksi jaettu opetus- ja testiaineistoon 70/30 jaolla satunnaisotantaa käyttäen. Jokaisella luokittelumallille satunnaisotanta oli tehty 100 kertaa. 95 %:n luottamusvälien käytettiin päätösrajojen tilastollisen merkitsevyyden arvioinnissa, mikä lisäsi tutkimuksen luotettavuutta. Vaihtoehtoinen päätösraja hyväksyttiin, jos tunnusluvun arvo oli parempi kuin  $P=0,5$  arvo ja vaihtoehtoisen päätösrajan 95%:n luottamusväli ei sisältänyt  $P=0,5$  tunnusluvun arvoa.

Tietyt aineiston ominaisuudet kuten, kuten epätasapaino, vaikuttivat tulosten validiteettiin ja tämä voi rajoittaa tulosten yleistettävyyttä muihin aineistoihin.

Jatkotutkimusehdotuksena voisi olla hyödyllistä tutkia menetelmiä, joilla voidaan parantaa luokittelutarkkuutta epätasapainoisten aineistojen luokittelussa. Olisi myös hyvä tutkia tarkemmin sitä, miten paljon aineiston epätasapainolla on vaikutusta luokittelukykyyn tai miten parantaa aineiston laatua.

## Lähteet

Alpaydin, E. 2021. Koneoppiminen. Helsinki: Terra Cognita Oy.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. New York: Springer Science+Business Media, LLC.

Burkov, A. 2019. The Hundred-Page Machine Learning Book. Andriy Burkov.

Burzykowski, T., Geubbelmans, M., Rousseau, A.-J. & Valkenborg, D. 2023. Validation of machine learning algorithms. American Journal of Orthodontics and Dentofacial Orthopedics, 164(2), 295–297. Saatavissa: <https://doi.org/10.1016/j.ajodo.2023.05.007>

Chicco, D. & Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21 (6). Viitattu 8.5.2023. Saatavissa <https://doi.org/10.1186/s12864-019-6413-7>

Chicco, D., Tötsch, N. & Jurman, G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 14 (13). Viitattu 8.5.2023. Saatavissa <https://doi.org/10.1186/s13040-021-00244-z>

Duodecim Käypä hoito. BDI-kysely. Viitattu 16.5.2023. Saatavissa <https://www.kaypa-hoito.fi/xmedia/pgr/BDI.pdf>

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. New York: Springer.

Kaakinen, M. & Ellonen, N. 2023. Logistinen regressio. Yhteiskuntatieteellinen tietoarkisto, Tampereen yliopisto. Viitattu 20.11.2024. Saatavissa <https://www.fsd.tuni.fi/fi/palvelut/metelmaopetus/kvanti/regressio/logistinen/>

Kämäräinen, J. 2023. Koneoppimisen perusteet. Tallinna: Gaudeamus Oy.

Kelleher, J. & Tierney, B. 2021. Datatiede. Helsinki: Terra Cognita Oy.

Kotu, V. & Deshpande, B. 2019. Data Science: Concepts and Practice. Second edition. Cambridge, MA: Morgan Kaufmann.

Murphy, K. P. 2012. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press.

Nummenmaa, L. 2021. Tilastotieteen käsikirja. Helsinki: Tammi.

Nummenmaa, L., Holopainen, M. & Pulkkinen, P. 2014. Tilastollisten menetelmien perusteet. Helsinki: Sanoma Pro Oy.

Panda, N. R., Kumar, P. J., Mohanty, J. N. & Bhuyan, R. 2022. A Review on Logistic Regression in Medical Research. National Journal of Community Medicine 13 (4), 265–270. Saatavissa <https://doi.org/10.55489/njcm.134202222>

Synlab. Tutkimusten tulostasot ja viitearvot. Viitattu 16.5.2023. Saatavissa <https://stage-www.synlab.fi/tietopankki/tutkimusten-tulostasot-ja-viitearvot/>

Tuominen, H. & Neittaanmäki, P. 2019. Tekoälyn perusteita ja sovelluksia. Viitattu 16.5.2023. Saatavissa <https://tim.jyu.fi/print/kurssit/tie/tiep1000/tekoalyn-sovellukset/kirja>

Uikkanen, E. 2024. Neuroverkon laskenta-algoritmeista. Viitattu 25.11.2024. Saatavissa <https://www.einouikkanen.fi/AI/Neuroverkon%20laskenta-algoritmeista.pdf>

Valmet. Viitattu 16.5.2023. Saatavissa <https://www.valmet.com/fi/valmet-yrityksena/valmet-lyhyesti/>

Valmet. Liiketoimintalinjat. Viitattu 16.5.2023. Saatavissa <https://www.valmet.com/fi/sijoittajat/valmet-sijoituskohteena/liiketoimintalinjat/>

Zheng, A. 2015. Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls. Sebastopol, CA, USA: O'Reilly Media