Ristomatti Paulin

Maastohiihtosuksen voiteen suositteleminen keliolosuhteen mukaan

Insinööri

Tieto- ja viestintätekniikka

Kevät 2024

KAMK • University of Applied Sciences

Tiivistelmä

Tekijä: Paulin Ristomatti

Työn nimi: Maastohiihtosuksen voiteen suositteleminen keliolosuhteen mukaan

Tutkintonimike: Insinööri (AMK), Tieto- ja viestintätekniikka

Asiasanat: hiihto, koneoppiminen, suksien voitelu

Tämän opinnäytetyön tarkoituksena oli yrittää kehittää koneoppimisalgoritmi, joka pystyisi suosittelemaan maastohiihdossa käytettäviä luistovoiteita keliolosuhteen perusteella. Opinnäytetyön toimeksiantajana toimi Jyväskylän yliopiston liikuntatieteellisen tiedekunnan liikuntateknologian Vuokatin yksikkö. Kehitettävän algoritmin tärkeimpänä tehtävänä on etenkin tehtyjen voidetestien voidetestitietokannan hyödyntäminen.

Hiihdossa suksen luisto- ja pito-ominaisuuksiin vaikuttaa useampi tekijä. Näitä ovat esimerkiksi painovoima, kitka, lumi, ilmanvastus, suksen voitelu ja sukseen tehty kuviointi, joita on avattu tarkemmin opinnäytetyön teoriaosiossa. Kitka ja ilmanvastus ovat etenkin hiihtäjän liikkumista estäviä voimia. Koska painovoima vaikuttaa aina suoraan alaspäin, se on alamäessä vauhtia lisäävä tekijä. Lumella tarkoitetaan sekä taivaalta satavaa lunta että jo maassa valmiiksi olevaa lunta, vaikka nämä ovat fysikaalisesti hyvin erilaisia. Suksen voitelulla ja kuvioinnilla saadaan parannettua suksen luistamista.

Suosittelualgoritmi pohjautuu koneoppimiseen. Koneoppiminen jaetaan yleensä ohjattuun ja ohjaamattomaan oppimiseen, ja lisäksi se voidaan jakaa myös vahvistusoppimiseen. Koska suosittelualgoritmi rakennettiin olemassa olevan aineiston pohjalta, käytettiin tässä opinnäytetyössä kehitetyissä koneoppimimalleissa ohjattua oppimista. Kehitystyössä testatut koneoppimismallit perustuvat päätöspuihin ja neuroverkkoihin.

Opinnäytetyössä kehitettyjen voiteiden suosittelualgoritmien ennustuskyky jäi toivuttua huonommaksi. Paras kehitetty suosittelu algoritmi, tarkkuuden perusteella, ylsi noin 20 % tarkkuuteen, jos sitä verrataan suosittelualgoritmiin, joka ennustaa testeissä käytetyintä voidetta. Tällaisen suosittelu algoritmin tarkkuus ylsi noin 13 % tarkkuuteen. Yksi syy sille, miksi voiteita voi olla hankala ennustaa on se, että useampi voide voi olla hyvä tietylle keliolosuhteelle, mutta tarkkuuteen perustuvalle luokittelijalle tämä on haastavampi määrittää. Toinen syy huonoille ennustustuloksille saattaa olla kehitetylle algoritmille hyödyllisen aineiston vähäinen määrä. Kolmantena syynä huonoille ennustustuloksille on aineiston- ja aineiston muuttujien laatu.

KAMK • University of Applied Sciences

Abstract

Author: Paulin Ristomatti

Title of the Publication: Recommending ski wax for cross-country skiing based on weather and ski track conditions

Degree Title: Bachelor of Engineering, Information and Communication Technologies

Keywords: nordic skiing, machine learning, nordic skiing waxing

The purpose of this thesis was to try to develop a machine learning algorithm that would be able to recommend ski waxes used in cross-country skiing based on the weather and ski tracks conditions. The Vuokatti unit of the sports technology of the Faculty of Sports Science of the University of Jyväskylä acts as the client of the thesis. The most important task of the developed algorithm is the utilization of the wax test database and to develop useful features to it.

In skiing, the gliding and grip properties of a ski are affected by several factors. These include, for example gravity, friction, snow, air resistance, ski wax and grinding as well as structuring done, which are explained in more detail in the theory section of the thesis. Friction and air resistance are especially the forces preventing the movement of the skier. Since gravity always acts directly downwards, it has a speed increasing factor when going downhill. Snow refers to both snow falling from the sky and snow already on the ground. Although these are physically quite different. Waxing and patterning the ski improves the gliding of the ski.

The recommendation algorithm is based on machine learning. Machine learning is usually divided into supervised and unsupervised learning, and it can also be divided into reinforcement learning. Since the recommendation algorithm is built on the existing data. Supervised learning is used in the machine learning models developed in this thesis. The machine learning models assessed in the development work were based on decision trees and neural networks.

The prediction ability of the wax recommendation algorithms developed in the thesis remained worse than it was hoped. The best developed recommendation algorithm, based on accuracy, reached about 20 % accuracy. If it is compared to a recommendation algorithm that recommends only the most used wax in the tests. The accuracy of this recommendation algorithm reached about 13 % accuracy. One of the reasons why waxes can be difficult to predict is that several waxes can be good for a given weather and ski tracks conditions, but for an accuracy-based classifier this is more challenging to determine. Another reason for the poor prediction results may be the amount of data useful for the developed algorithm. And thirdly the quality of the data and data variables.

Sisällys

Symboliluettelo

DBSCAN	Density-based spatial clustering of applications with noise, Kohinaa sisältävien sovellusten tiheyteen perustuva spatiaalinen klusterointi
FIS	Fédération Internationale de Ski, kansainvälinen hiihtoliitto
GPS	Global Positioning System, maailmanlaajuinen paikallistamisjärjestelmä
ID	Identifier, yksilöllinen tunniste
IQR	Interquartile range, Kvartiiliväli, Kuvaa havaintoarvojen hajaantuneisuutta
MICE	Multiple Imputation by Chained Equations, Moninkertainen imputointi ketjutetuilla yhtälöillä
MLP	Multi-Layer Perceptron, Monikerroksinen perseptroniverkko
OHE	One-Hot Encoding, Koodataan muuttujan kaikki uniikit arvot binääri sarakkeiksi
PCA	Principal components analysis, pääkomponenttianalyysi
PFAS	Per- ja polyfluoratut alkyyliyhdisteet
PFOA	Perfluorioktaanihappo
Q1	Alakvartiili
Q3	Yläkvartiili
ReLU	Rectified Linear Unit, Oikaistu lineaarinen yksikkö, epälineaarinen aktivointifunktio
TF	TensorFlow, koneoppimisohjelmistokirjasto

1 Johdanto

Opinnäytetyö käsittelee perinteisen ja luistelu hiihtotavan suksien rakennetta ja sitä, mitkä asiat vaikuttavat suksien luisto-ominaisuuksiin. Näiden pohjalta pyritään myöhemmin rakentamaan koneoppimismalli suosittelemaan suksien voitelua keliolosuhteiden mukaisesti. Opinnäytetyö tehdään osana toimeksiantajan, Jyväskylän yliopiston, suksihuoltohanketta (Suksihuollon kehityshanke OVK Vuokatti-Rukan lajeissa, EAKR 2023–2025). Hankkeessa on tavoitteena kehittää suomalaista lumilajien välinehuoltoa, testaamista ja välinehuollon yhteistyötä sekä koulutusta OVK Vuokatti-Rukan lajeissa, joita ovat pikkusuksien lajit (ampumahiihto, maastohiihto, yhdistetty ja mäkihyppy) sekä rinnelajit (alppihiihto, lumilauta, freeski ja kumparelasku). [1.]

Maastohiihdon suksen rakenne muodostuu nykyään etenkin kilpahiihdossa kennorungosta ja pohjassa käytetystä polyeteenimuoviyhdistelmästä, joka on sintrattu muovipohja. Pohjamateriaalin valinnalla vaikutetaan pohjan huokoisuuteen ja täten myös sen voideltavuuteen. Kennorunkoinen suksi on myös kevyempi, elastisempi ja aggressiivisempi verrattuna puu- ja vaahtorakenteiseen ytimeen. [2.]

Maastohiihdon suksen pito- ja luisto-ominaisuuksiin vaikuttaa moni asia. Näistä etenemistä haittaavia on etenkin kitka- sekä painovoima ja ilmanvastus tietyissä olosuhteissa, koska painovoimasta ja tuulesta voi myös olla hyötyä esimerkiksi alamäkeen laskettaessa. Lisäksi lumen lämpötilalla, olomuodolla ja kosteudella on suuri vaikutus suksen luistamiseen ja suksissa käytettävien voiteluaineiden valintaan.

Jotta voiteiden valintaa voidaan suositella automaattisesti, tarvitaan tekoälyä. Tekoäly on laaja käsite ja sillä tarkoitetaan tässä opinnäytetyössä lähinnä koneoppimista. Koneoppiminen jaetaan yleensä kolmeen algoritmityyppiin, jotka ovat ohjattu oppiminen, vahvistusoppiminen ja ohjaamaton oppiminen. Ohjatussa oppimisessa koneoppimismallin opettamiseen käytettävästä opetusdatasta tiedetään ennalta haluttu ulostulo ja tällä tavoin mallia opetetaan antamaan tämä tulos. Vahvistusoppimisessa kyse on agentin sopivien toimien toteuttamisesta palkitsemisen maksimoimiseksi tietyssä tilanteessa. Vahvistusoppimisessa agentti oppii siis tekemään päätöksiä, joista se palkitaan kaikista parhaiten. Ohjaamattomassa oppimisessa ei tiedetä ennalta haluttua lopputulosta. Klusterointi on yksi yleisin ohjaamattomassa oppimisessa käytetty menetelmä. Klusteroinnissa aineisto jaetaan niin, että kunkin luokan alkiot muistuttavat toisiaan enemmän

kuin muiden luokkien alkioita [3]. Tässä opinnäytetyössä koneoppimisalgoritmi tyypeistä keskitytään lähinnä ohjaamattomaan oppimiseen ja ohjattuun oppimiseen, joista jälkimmäistä hyödynnetään satunnaismetsien ja neuroverkkojen tyyppisten koneoppimismallien avulla.

2 Maastohiihto

Tässä luvussa käsitellään maastohiihdossa käytettyjen perinteisen ja vapaan hiihtotavan suksia. Vaikka opinnäytetyön aineistossa on käytetty vain vapaan hiihtotavan suksia, on perinteisen hiihtotavan suksien rakeenteen ymmärtämisestä höytyä myös vapaan hiihtotavan suksien rakenteen ymmärtämisessä.

2.1 Perinteisen hiihtotavan sukset

Veli Kolehmaisen tutkielmassa viitatun vuonna 1985 julkaistun Heikki Kantolan Sykettä ladulle -kirjan mukaan perinteisen hiihtotavan suksen toiminnalliset alueet voidaan jakaa suksen etu- ja takaosassa oleviin liukupainealueisiin ja suksen keskellä olevaan pitoalueeseen. Suksen tärkeimmät toiminnalliset osat ovat liukupainealueet. Muuttuvat olosuhteet ja hiihtosuorituksen vaiheet asettavat suksen rakenteelle ja toiminnalle erilaisia vaatimuksia. [2.] Nykyaikainen maastohiihtosuksien rakenne on pääsääntöisesti aina samanlainen. Pinta sekä pohja suksessa on muovia ja runko koostuu joko puu-, vaahto- tai kennorakenteisesta ytimestä. Kennorakenteisella ytimellä valmistetut sukset ovat kevyempiä, elastisempia ja aggressiiviisempia. [4., s 33]

Tämän vuoksi yleensä kilpahiihtäjien sukset on valmistettu kennorunkoon. Puuytimen omaavat sukset ovat painavampia sekä rauhallisempia ja tätä kautta monesti helpommin hiihdettäviä. Vaahtorakenteisella ytimellä olevat sukset ovat ominaisuuksiltaan näiden edellä mainittujen suksien väliltä. Suksien pinta- ja pohjamateriaaleissa käytetään erilaisia polyeteenimuoviyhdisteitä riippuen suksen hintaluokasta. Halvemmissa suksissa käytetään ekstrahoitua muovipohjaa, kun taas kalliimman hintaluokan suksissa käytetään sintrattua muovipohjaa. Pohjamateriaalin valinnalla voidaan vaikuttaa pohjan huokoisuuteen ja täten myös sen voideltavuuteen. [4., s 33]

2.2 Vapaan hiihtotavan sukset

Vapaan suksia on tutkittu vähän. Suksen keulan rakenne, paino ja muoto vaikuttavat kuitenkin tasapainopisteen sijoittumiseen. Takavuosien suurimmat innovaatiot olivatkin kärkien huomattava keventyminen ja suksen tasapainopisteen siirtäminen taaksepäin. Tasapainopisteen paikka taas vaikuttaa osaltaan etu- ja takapainealueiden sijoitteluun. Vapaan suksien välillä on edelleen

selkeitä eroja hiihtotuntumassa ja tasapainopisteen sijainnissa. Myös suksien kärkien käyttäytyminen eroaa. Suksen keulan on annettava riittävä tuki ponnistukseen myös kovalla alustalla, mutta toisaalta kärki ei saa leikata kiinni lumeen ylämäkeen luistellessa. [5.]

Hiihtotekniikka vaikuttaa suksen hiihdettävyyteen. Jos hiihtäjä ei osaa viedä vartalon painoa eteen jalkaterän muodostaman tukialueen päälle, takapainoiseksi rakennettu suksi ei toimi hänelle siten kuin se on suunniteltu. Keulan toiminta ja esimerkiksi sopiva kiertojäykkyys vaikuttavat merkittävästi, miten suksi toimii ylämäessä ja palautuu hiihtäjän alle potkun jälkeen. Tasapainopiste vaikuttaa luistelusuksen käyttäytymiseen, koska painealueet ja kaaren korkein kohta määräytyvät osittain myös tasapainopisteen kautta. Värikkään transparenttipohjan kova materiaali hylkii likaa mustaa pohjaa paremmin ja toimii myös roskaisella ja kostealla kevätlumella. [5.]

3 Suksen luisto-ominaisuuksiin vaikuttavat tekijät

Tässä opinnäytetyön luvussa keskitytään maastohiihdon suksen luisto-ominaisuuksiin vaikuttaviin tekijöihin. Luisto-ominaisuuksiin vaikuttavia tekijöitä on muun muassa painovoima, kitka, lumi, ilmanvastus, suksen voitelu ja suksen konekuviointi sekä suksen käsikuviointi.

3.1 Painovoima

Voiman kaava F määritetään kaavalla. [6].

$$F = ma \qquad (1)$$

missä m on massa ja a on kiihtyvyys.

Painovoima vaikuttaa hiihtäjän etenemiseen. Riippumatta hiihtoalustan kaltevuudesta painovoima vaikuttaa aina suoraan alaspäin. Mäkisessä maastossa painovoiman aiheuttamat voimat vaikuttavat hiihtäjään joko vauhtia lisäävänä tai hidastavana tekijänä. Voima lasketaan kaavasta (1). Painovoimaan vaikuttava kiihtyvyys on putoamiskiihtyvyys, joka on noin 9,81 m/s^2 lähellä maan pintaa ja sitä merkitään SI-järjestelmässä g-kirjaimella.

3.2 Kitka

Kitkavoima F_f määritetään kaavalla. [7].

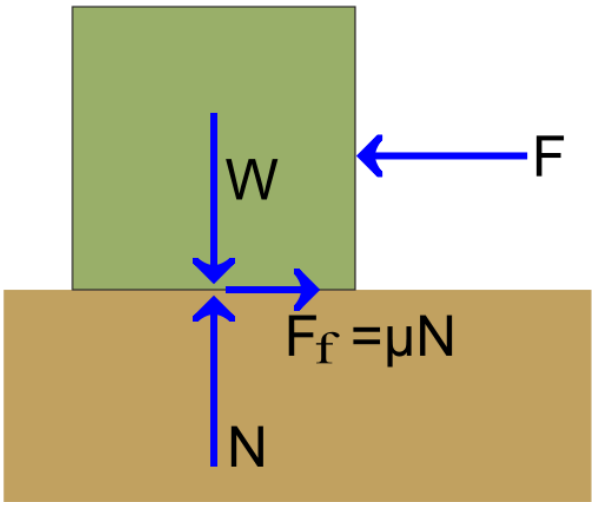
$$F_f = \mu N \qquad (2)$$

Missä μ on kitkakerroin ja N on pinnan tukivoima. Pinnan tukivoima N kuvassa 1. on yhtä suuri kuin voima W, joka saadaan laskemalla kaavalla 1.

Kitka on voima, joka vastustaa kahden toisiaan koskevan kappaleen välillä ilmenevää liikettä tai tällaisen liikkeen alkamista. Kuvassa 1. on havainnoitu kitka voimaa ja siihen vaikuttavia voimia. Kitkavoima johtuu siitä, että alustan ja sitä vasten liukuvan esineen välisten pienten epätasaisuuksien reunat on murrettava, jotta esine voisi liukua toista vasten. Kitkavoima vaikuttaa siis aina

liikutettavan kappaleen liikkeen suuntaa vastaan. [7]. Kitka yleisesti jaetaan lepo- ja liikekitkaan, joista lepokitka vaikuttaa kappaleiden välisessä liikkeellelähdössä ja jälkimmäinen liikkeen jatkuessa. Kitkan vaikutukseen voidaan vaikuttaa materiaaliparien valinnalla, kappaleiden pinnan muokkaamisella ja käyttämällä voiteluaineita. [2.]

Suksen pysty- ja vaakavoimia pystytään mittaamaan esimerkiksi Jyväskylän yliopiston liikuntateknologian yksikön rakentamalla suksen liikutuslaitteella [8].



Kuva 1. Kitkavoima F_f kaavasta (2) vastustaa voimaa F [9]

3.3 Lumi

Lumi ei koskaan voi olla lämpötilaltaan lämpimämpää kuin 0 ^{0}C, koska vesi alkaa jäätymään alle nollan celsiusasteen lämpötilassa. Vapaan veden määrä lumessa kuitenkin vaihtelee eri lämpötilassa. Lumella tarkoitetaan sekä taivaalta satavaa että jo maassa olevaa lunta. Fysikaalisesti nämä kaksi eroavat toisistaan kuitenkin huomattavasti. Otollisissa olosuhteissa lumikide jatkaa kasvuaan härmistymisen vaikutuksesta, jolloin niistä muodostuu lumihiutaleita. Lumihiutaleiden massa kasvaa lopulta niin suureksi, että maan vetovoima vetää niitä alaspäin. Ilman lämpötilan mukaan lumihiutale voi kasvattaa tai pienentää kokoaan vielä maahan putoamisen aikana. [2.]

Maahan laskeutuneet lumikiteet muuttavat muotoaan jatkuvan termomekaanisten prosessien eli metamorfoosien avulla. Metamorfoosiin vaikuttaa lumikiteitä ympäröivät olosuhteet, kuten ilman lämpötila, auringon säteily, lumi- tai vesisade, tuuli, maaperän geoterminen lämpö ja painovoima. Vaikuttavan metamorfoosin kautta lumen koostumus muuttuu ja maata peittävä lumipeite kerrostuu, näistä useista eri paksuisista ja muotoisista lumikidekerroksista. Lumipeitteen tiheys, huokoisuus, lämmönjohtavuus, heijastuskyky ja kokoonpuristuvuus on siis koko talven ajan jatkuvassa muutoksessa metamorfoosien takia. [2.]

Lumen termodynaamiset ominaisuudet, kuten sen lämmönjohtavuus, vaikuttavat myös suksia vasten muodostuvan kitkakertoimen suuruuteen. Lämpötilan laskiessa lämmönjohtavuus heikkenee, mikä vaikuttaa lumen ja suksen välisen ohuen vesikerroksen muodostumiseen. Kitkan vaikutuksesta suksen pohjan lämpötila taas nousee yhdestä neljään celsiusastetta riippuen liikenopeudesta sekä vaikutuspisteestä suksen pohjassa. Lämmönjohtavuus ja auringon säteilyn vaikutus onkin syytä huomioida suksien pohjamateriaaleja ja voiteita valmistettaessa. [2.]

Lumipeitteen kovuus kasvaa lineaarisesti lämpötilan laskiessa. Suksen luiston kannalta olisikin oleellista, että suksen pohjan sekä voiteiden kovuuden tulisi vastata lumen kovuutta. Mitä kovempaa lumi on, sitä parempi on myös suksen luisto. Lumen ollessa vanhaa lunta merkittäväksi luistoa heikentäväksi tekijäksi nousee lumen vesipitoisuus. Valitsevalla olosuhteella sekä lumen olomuodolla on merkittävä vaikutus suksen luistolle ja se on syytä ottaa huomioon tarkempaa suksi- ja voidevalintaa tehtäessä. [2.]

3.4 Ilmanvastus

Newtonin vastuslaki määritetään kaavalla. [10].

$$F_v = \frac{1}{2}\rho v^2 A C_v \quad (3)$$

Missä ρ on ilman tiheys, v on kappaleen nopeus, A on kappaleen poikkipinta-ala ja C_v on ilmanvastuskerroin.

Ilmanvastuksella tarkoitetaan kappaleen liikettä vastustavaa voimaa, joka aiheutuu kappaleen pinnan ja ilman hiukkasten välisestä vuorovaikutuksesta. Newtonin vastuslaista (3) voimme huomata ilmanvastuksen kasvavan nopeuden neliöön. Tämän takia sillä on merkittävä vaikutus, mitä

suuremmaksi nopeus kasvaa. Ilmanvastukseen vaikuttavat kappaleen koko, muoto sekä ilmanpaine ja tuuli. Tuulen vaikutus maastohiihtäjään nähden voi olla joko vauhtia lisäävä tai vähentävä tekijä. Kovemmassa vastatuulessa maastohiihtäjä voi joutua muuttamaan hiihtoasentoaan tai tekniikkaansa. Hiihtäjä voi pienentää ilmanvastuksen vaikutusta alamäessä laskemalla etukumarassa, vaatetuksellaan tai laskemalla edellä hiihtävän perässä, jolloin edessä hiihtävälle hiihtäjälle voi olla jopa hieman hyötyä turbulenssin pienentymisen takia. [2.]

3.5 Voitelu

Suksen voitelussa on tärkeää, että suksen pohja on liasta puhdas ennen kuin siihen levittää uutta voidetta. Uuden luistovoiteen levitys likaiseen pohjaan kiinnittää kertyneen lian syvemmälle pohjaan ja heikentää luistoa. [11]. Hyvän luiston ansiosta suksi liukuu kevyesti eteenpäin. Kuntohiihtäjä pärjää yleensä hyvin helppokäyttöisillä ja nopeilla pikavoiteilla. Pikavoitelu on nopea tehdä, mutta kuluu myös pois nopeammin kuin perusteellisesti tehty voitelu. Tämän takia suksen pohjustusvoide on tärkeä laitattaa aika ajoin. Pikavoitelussa voide levitetään sukseen pullon omalla sienellä, annetaan hetki kuivua ja lopuksi kiillotetaan liinalla. Luistelusuksien voitelu eroaa perinteisten suksien voitelusta, sillä luistelusuksissa ei käytetä ollenkaan pitovoidetta. Luistelusuksien voitelussa ehdottomasti tärkein asia on pohjan puhtaus. [12.]

Fluorivoiteet ovat ylivertaisia hylkimään vettä ja likaa sekä säilyttämään suksen luisto-ominaisuuden. 1980-luvun lopulta tähän päivään kaikki jaossa olleet olympia- ja MM-mitalit on voitettu fluorivoidelluilla suksilla. Fluorin käyttöä rajoittavaa lainsäädäntöä on tiukennettu Euroopan unionissa lähivuosina. Mitä tulee hiihdossa käytettäviin fluorivoiteisiin, nykyinen laki asettaa raja-arvon ainoastaan fluorivoiteissa olevalle perfluorioktaanihapolle eli PFOA:lle ja sen esiasteille. Aiemmin PFOA:ta ilmeni raja-arvot ylittäviä määriä vahvoissa fluorivoiteissa kuten C8:ssa, jonka nimessä C kuvaa hiiltä ja numero 8 hiiliatomien määrää. Lyhyesti ilmaistuna: Mitä isompi hiiliatomien lukumäärä on, sitä vahvempaa fluoritavara on. Heinäkuussa 2020 voimaan tulleen EU-lain myötä markkinoille on kuitenkin tullut puhdistettuja, PFOA:lle asetetut raja-arvot alittavia C8-ketjuisia fluorivoiteita. EU:ssa on halua kiristää eritoten per- ja polyfluorattujen alkyyliyhdisteiden eli PFAS-yhdisteiden käyttöä. Edellä mainittu PFOA kuuluu PFAS-yhdisteisiin. Voideteollisuuden ympäristökuormasta on vähän tutkimusta pitkälti sen vuoksi, että alan volyymi on promilleluokkaa verrattuna suuriin PFAS-päästölähteisiin, muun muassa tekstiiliteollisuuteen. [13.]

Kansainvälinen Hiihtoliitto (FIS) on ilmoittanut, että fluorivoide kielto tulee voimaan talvikaudelle 2023/2024. FIS on edelleen sitoutunut kieltämään suksien valmistuksessa käytettävät fluorituotteet, kun otetaan huomioon fluorivahoihin liittyvät terveysriskit ja ympäristöongelmat. FIS on kehittänyt yhteistyössä yhdysvaltalaisen Brukerin kanssa ALPHA II -testausmenetelmää ja samalla tiivistänyt yhteistyötä kansainvälisen ampumahiihtoliiton (IBU) kanssa yhteisen työryhmän kautta. FIS on talvella 2022/2023 käyttänyt laitteen laajan testauksen ja testauksen suorittavien toimihenkilöiden kouluttamiseen, jotta laite antaisi luotettavat tulokset fluorittomien kilpailujen varmistamiseksi. FIS suorittaa testejä varmistaakseen, että sukset ovat fluorittomia huipputason tapahtumissa, mukaan lukien FIS:n hiihdon maailmanmestaruuskilpailut, FIS:n maailmancup-kilpailut ja muut suuret tapahtumat, kuten FIS:n nuorten hiihdon maailmanmestaruuskilpailut. Testaus alemman tason tapahtumissa suoritetaan satunnaisesti sen varmistamiseksi, että myös nämä tapahtumat ovat säänneltyjä. [14.]

Suomen hiihtoliiton mukaan kansainvälisten lajiliittojen suositus fluorikiellosta ei astu voimaan kansallisella tasolla vielä kaudelle 2023/2024. Hiihtoliiton perusteina linjaukselle ovat aikataulu sekä rajalliset resurssit valvonnan toteuttamiseksi. Suomessa tullaan kuitenkin noudattamaan fluorien käyttöä koskevaa EU-säännöstä ja valmistelemaan kansallista fluorikieltoa. EU-säännöstön tavoitteena on vähentää fluorin haitallisten aineiden vaikutuksia terveyteen ja ympäristöön. Kansallisella tasolla valmistaudutaan fluorien käytön kieltämiseen. Tavoitteena on kiellon täysimääräinen voimaantulo ja valvonnan aloittaminen kauden 2024/2025 aikana. Fluorikieltoa edistää kansallinen työryhmä, jossa asiaa työstävät Hiihtoliiton sekä Ampumahiihtoliiton edustajat yhdessä kilpailu- ja tuomaritoiminnan edustajiston kanssa. Asiantuntijat rakentavat kansallista säännöstöä kansainvälisten linjausten mukaan. [15.]

3.6 Kuviointi

Hiihtäessä suksien muodostama paine saa lumen pintakerroksen sulamaan pohjaa vasten. Sulaminen periaatteessa pienentää kitkaa, mutta vesikerros kahden tasaisen pinnan välillä saa myös aikaan liisteriefektin, jota voi verrata kahden märän lasilevyn liu'uttamiseen vastakkain. Kuviokone työstää suksen pohjan voitelukerrokseen ohuen kuvion, joka lisää ilmavirran pyörteilyä suksen ja lumen välillä. Välissä oleva ilma vähentää vastusta ja suksi luistaa pidemmälle. Sama asia tapahtuisi myös lasilevyille, jos toiseen niistä tehtäisiin uria. Mitä kosteampi latu, sitä suurempi

etu kuvioinnilla yleensä saavutetaan. Kuvioiden karheus kasvaa myös kelin lämmetessä. Ero lähtötilanteeseen on saatu vielä selvemmäksi käyttämällä kullekin kelille optimoitua tietynmuotoista terää. Kuvioinnin vaikutus luistoon on nykysuksilla niin selvä, ettei kilpavoiteluiden tekeminen ilman kuviointia ole enää mielekästä. [16].

Käsikuvioinnilla tarkoitetaan kuviointia, joka ajetaan suksen pohjaan voitelun jälkeen. Kuvio muodostuu siis voitelukerrokseen ja se kuluu pois voiteen mukana. Selkeillä kuvioilla maksimoidaan ilmavirran pyörteily suksen pohjassa. Lumen ominaisuudet muuttuvat lämpötilan mukana, minkä vuoksi myös paras kuviointi muuttuu kelien mukaan. Karkeasti kuviot voidaan jakaa kolmeen eri tyyppiin suoriin kuvioihin, havukuvioihin ja lineaarisiin kuvioihin. Eri kuviotyypit on tarkoitettu erilaisiin keliolosuhteisiin. Kuvioita voidaan tehdä myös suksien pohjauriin, mikä auttaa entisestään poistamaan kosteutta suksen ja lumen välistä. [17.]

Konekuviointi tarkoittaa kivihionta koneella tehtävää hiontaa suksen pohjaan. Pääkomponentit kivihiontalaitteessa on kivi ja timantti. Timantilla uurretaan kiven pintaan kiven pyöriessä. Liikuttamalla timanttia poikittaissuunnassa pinnan yli kiven pyöriessä saa aikaan uria kiven pintaan. Hionnan aikana kivipinnan urakuvio tuottaa ja päättää suksien pohjan rakenteen. Suksen pohjarakenteen karheus vaihtelee timantin poikittaisnopeudesta kiven pinnan yli. Kiven pinnan mineraalit tylsistyvät suksien hionnan aikana. Kun kymmeniä suksi pareja on hiottu, kivi menettää kyvyn muodostaa teräviä uria suksenpohjaan. Kiven pinnan kuvio on tämän jälkeen rakennettava uudelleen. Timantti myös kuluu hionnan aikana ja se on vaihdettava aika ajoin. [18.]

4 Tekoäly

Tässä opinnäytetyön luvussa esitellään tekoälyä. Tässä opinnäytetyössä tekoälyllä tarkoitetaan lähinnä koneoppimista. Koneoppimisessa oppimismenetelmiä on ohjattu ja ohjaamaton oppiminen sekä vahvistusoppiminen, mihinkä ei perehdytä sen tarkemmin tässä opinnäytetyössä. Opinnäytetyössä kehitetyt ennustusalgoritmit kehitetään ohjatun oppimisen avulla.

4.1 Ohjattu oppiminen

Ohjattu koneoppiminen on nimensä mukaisesti ihmisten ohjaamaa oppimista. Ohjatussa oppimisessa tarkoituksena on luoda malli, joka ennustaa vastemuuttujan arvoja selittävien muuttujien avulla mahdollisimman hyvin. Koneoppimisessa selittäviä muuttujia kutsutaan usein ominaisuuksiksi tai piirteiksi (eng. features) ja vastemuuttujaa kutsutaan kohdemuuttujaksi (eng. target). Tarkoituksena ohjatussa oppimisessa on siis löytää funktio, joka kuvaa hyvin syötemuuttujien ja vastemuuttujien välistä suhdetta. [3] [19].

4.2 Ohjaamaton oppiminen

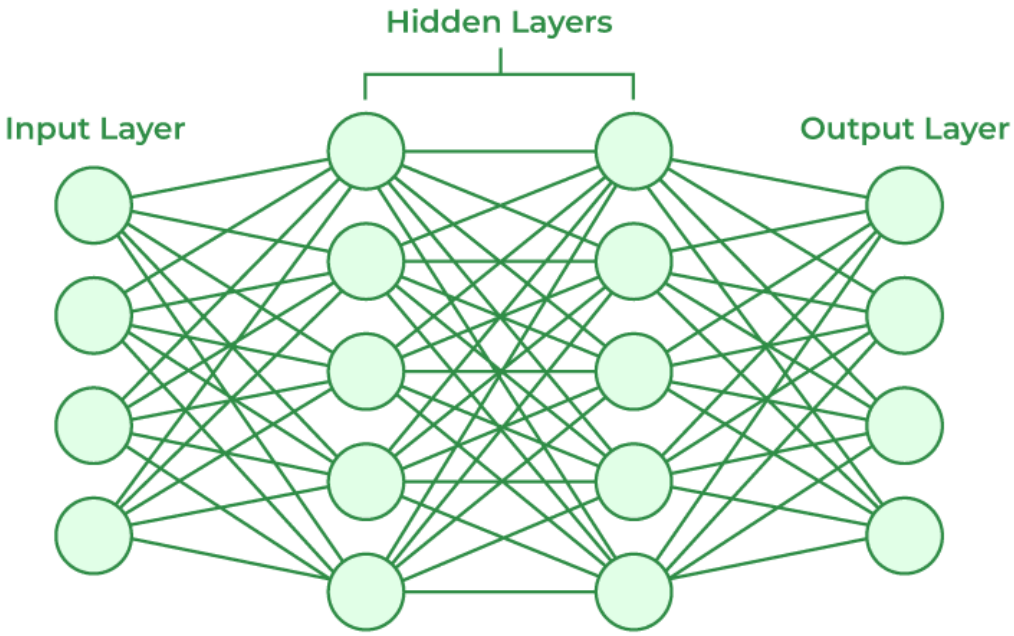
Ohjaamatonta oppimista käytetään, kun käytettävissä aineistossa ei ole syöte-vaste-pareja (eng. labeled data). Toisin sanoen sitä käytetään, kun aineistossa ei ole kohdemuuttujaa, jonka arvot ovat tiedossa. Ohjaamattoman oppimisen avulla pyritään löytämään yhteyksiä muuttujien välillä niin, että aineiston havainnot voidaan jakaa jonkinlaisiin ryhmiin. Kun ohjatun oppimisen tavoitteena on luokitella dataa halutulla tavalla, niin ohjaamattoman oppimisen tavoitteena on etsiä piilossa olevia riippuvuuksia datasta. [3] [19].

4.3 Koneoppimismallit

Tässä opinnäytetyön alaluvussa esitellään kehitettäviä koneoppimismalleja, jotka ovat syvä neuroverkko ja satunnaismetsä.

4.3.1 Syvät neuroverkot

Kuvassa 2. on esitelty neuroverkon rakennetta.

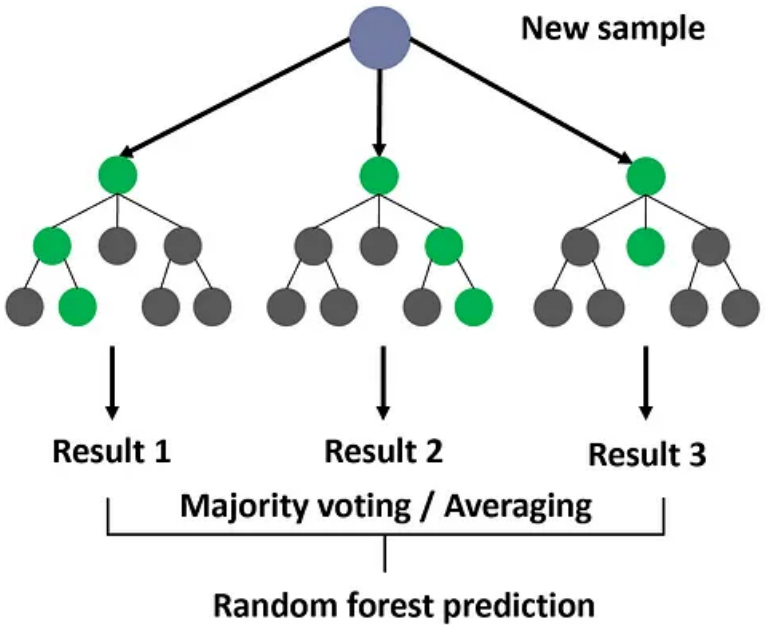


Kuva 2. Neuroverkon rakenteesta. [20].

Neuroverkko on algoritmi, joka perustuu ihmisten aivojen toimintatapaan. Se koostuu keinotekoisista neuroneista, jotka perustuvat biologisiin neuroneihin. Neuroverkkojen tarkoituksena on kopioida aivojen toimintaa, missä neuronit keskustelevat toistensa kanssa lyhyillä sähkösignaaleilla. Aivoissa jokainen neuroni saa signaaleja tuhansista muista neuroneista niin, että jokaisella signaalilla on oma painoarvo. Kun nämä signaalit saapuvat neuroniin, ne yhdistyvät jonkinlaisen kaavan mukaan. Kun tämän kaavan tulos ylittää jonkin rajan, neuroni aktivoituu (aktivointi funktio). Tällaista neuronien välistä keskustelua yritetään saavuttaa neuroverkoilla. Myötäkytkentäinen (eng. forward feed) neuroverkko on yksi suosituimmista neuroverkoista, missä syöttö kulkee syöttökerroksesta aina seuraavan kerrokseen, josta se lopuksi saapuu ulostulokerrokseen. Jokaisessa neuroverkossa pitää olla syöttökerros ja ulostulokerros. Lisäksi neuroverkossa voi myös olla piilokerroksia. Se mikä neuroverkosta tekee syvän, on piilokerroksien määrä. [21.]

4.3.2 Satunnaismetsät

Kuvassa 3. on esitelty satunnaismetsän rakennetta.



Kuva 3. Satunnaismetsän rakenteesta [22].

Satunnaismetsä-koneoppimismenetelmä koostuu useasta päätöspuusta. Kukin päätöspuu muodostetaan käyttämällä jotain algoritmia. Satunnaismetsän ennustus muodostuu päätöspuiden perusteella siten, mitä suurin osa päätöspuista on ennustanut [23].

Päätöspuut ovat olennainen osa satunnaismetsän toimimisessa. Päätöspuut muodostuvat noodeista ja niistä yhdistävistä ”poluista”. Päätöspuussa lähdetään liikkeelle aloitusnoodista ja päädytään lopetusnoodiin riippuen ratkaistavasta ongelmasta. Päätöspuun luokittelu on rekursiivinen prosessi, missä aikaisempien noodien päätökset vaikuttavat tämänhetkiseen noodiin. Vaikka päätöspuut ovat helppokäyttöisiä, ne pystyvät siitä huolimatta käsittelemään monityyppistä dataa. Kuitenkin datamäärän kasvaessa päätöspuiden muodostaminen hidastuu. Päätöspuiden kaksi yleisintä ongelmaa on ylisovittuminen liian syvien päätöspuiden takia ja se, että tietyn noodin sijainti puussa voi vaikuttaa ennustettuun tulokseen merkittävästi. Nämä ongelmat satunnaismetsä pyrkii ratkaisemaan satunnaisesti muodostetuilla päätöspuilla. [24.]

4.4 Arviointimetriikat

Luistovoidetta ennustaessa tärkein algoritmin toimivuutta arvioiva mittari ei ole välttämättä oikein ennustettujen luistovoiteiden määrä eli tarkkuus, koska kaksi samantyylistä voidetta voi olla yhtä hyviä. Ennustuksen tyyppi on myös monesta luokasta ennustaminen (eng. Multi-class classification), jota ei kannata sekoittaa monen luokan ennustamiseen (eng. Multi-label classification) ja aineistossa luokat ovat myös epätasapainoisuudessa. Se tarkoittaa esimerkiksi sitä, että luokkaa A on aineistossa enemmän tai vähemmän kuin luokkia B tai C. Näiden syiden takia monesta luokasta ennustaessa olisi mielekästä käyttää muitakin algoritmin toimivuutta mittaavia metriikoita. Näitä muita algoritmin suorituskykyä mittaavia mittareita on ainakin sisäinen tarkkuus (eng. Precision), herkkyys (eng. recall), F1-arvo.

Tietyn luokan sisäinen tarkkuus lasketaan kaavalla [25]

$$Sisäinen\ tarkkuus_{luokka\ A} = \frac{TP_{luokka\ A}}{TP_{luokka\ A} + FP_{luokka\ A}} \qquad (5)$$

, missä TP tarkoittaa true positive eli oikeat positiiviset ja FP tarkoittaa false positive eli väärät positiiviset. Sisäinen tarkkuus mittaa mallin kykyä tunnistaa tietyn luokan esiintymät oikein. Esimerkki väärästä positiivisesta on, jos luokan A:n arvo on ennustettu luokaksi B [25].

Tietyn luokan herkkyys lasketaan kaavalla [25],

$$Herkkyys_{luokka\ A} = \frac{TP_{luokka\ A}}{TP_{luokka\ A} + FN_{luokka\ A}} \qquad (6)$$

, missä TP on oikeat positiiviset ja FN on false negative eli väärät negatiiviset. Herkkyys mittaa mallin kykyä tunnistaa tietyn luokan kaikki esiintymiset. Esimerkki väärästä negatiivisesta on, jos luokkaa A tarkasteltaessa luokaksi A on ennustettu luokan B arvo. [25]

Tosin sisäisen tarkkuuden ja herkkyyden laskeminen jokaiselle luokalle tarkoittaisi sitä, että metriikoita olisi kaksi kertaa niin paljon kuin aineistossa on uniikkeja luokkia. Tämän haasteen välttämäseksi sisäiset tarkkuudet ja herkkyydet voidaan laskea keskiarvoistamalla ne luokkien kesken. Keskiarvon laskemiseen kaksi yleisintä tapaa on mikro- ja makrokeskiarvo sisäiselle tarkkuudelle ja herkkyydelle.

Sisäinen tarkkuus ja herkkyys makrokeskiarvolla lasketaan kaavoilla

$$Sisäinen\ tarkkuus_{makrokeskiarvo} = \frac{Sisäinen\ tarkkuus_{luokka\ A} + \cdots + Sisäinen\ tarkkuus_{luokka\ N}}{N} \qquad (7)$$

$$Herkkyys_{makrokeskiarvo} = \frac{Herkkyys_{luokka\,A}+\cdots+Herkkyys_{luokka\,N}}{N} \quad (8)$$

Missä jokaisen yksittäisen luokan sisäiset tarkkuudet ja herkkyydet lasketaan yhteen ja jaetaan luokkien yhteismäärällä. Makrokeskiarvo antaa jokaiselle luokalle saman painoarvon, mikä on hyödyllistä silloin, kun kaikki aineiston luokat ovat samanarvoisia. [25].

Sisäinen tarkkuus ja herkkyys mikrokeskiarvolla lasketaan kaavoilla [25]

$$Sisäinen\ tarkkuus_{mikrokeskiarvo} = \frac{TP_A+\cdots+TP_N}{TP_A+FP_A+\cdots+TP_N+FP_N} \quad (9)$$

$$Herkkyys_{mikrokeskiarvo} = \frac{TP_A+\cdots+TP_N}{TP_A+FN_A+\cdots+TP_N+FN_N} \quad (10)$$

Missä TP_N on yksittäisen luokan oikeat positiiviset, FP_N on väärät positiiviset ja FN_N on väärät negatiiviset. Molemmat mikrokeskiarvot sisäinen tarkkuus ja herkkyys antavat saman tuloksen, koska luokan väärä positiivinen on toisessa luokassa väärä negatiivinen. Molemmat metriikat antavat myös saman tuloksen kuin tarkkuus, koska mittareissa jaetaan kaikki oikeat ennustukset kaikilla ennustuksilla, mikä on myös tarkkuuden määritelmä. [26].

F1-arvo binäärisessä luokittelussa lasketaan kaavalla [27]

$$F_1 = 2\frac{Sisäinen\ tarkkuus \cdot Herkkyys}{Sisäinen\ tarkkuus+Herkkyys} \quad (11)$$

F1-arvo on sisäisen tarkkuuden ja herkkyyden harmoninen keskiarvo, jota voidaan käyttää näiden mittareiden tasapainon löytämiseksi. Suurempi mittarin arvo kertoo mallin paremmasta suorituskyvystä. Moniluokkaluokittelussa voidaan käyttää myös F1-arvoa, mutta tämän laskemiseen on kaksi eri tapaa, jotka ei välttämättä anna samaa tulosta. F1-arvon voi joko laskea F1-arvojen keskiarvosta tai keskiarvoistetulla F1-arvolla. Keskiarvoistettu F1-arvo lasketaan kaavalla, missä aritmeettinen keskiarvo lasketaan harmonisista keskiarvoista. [27]

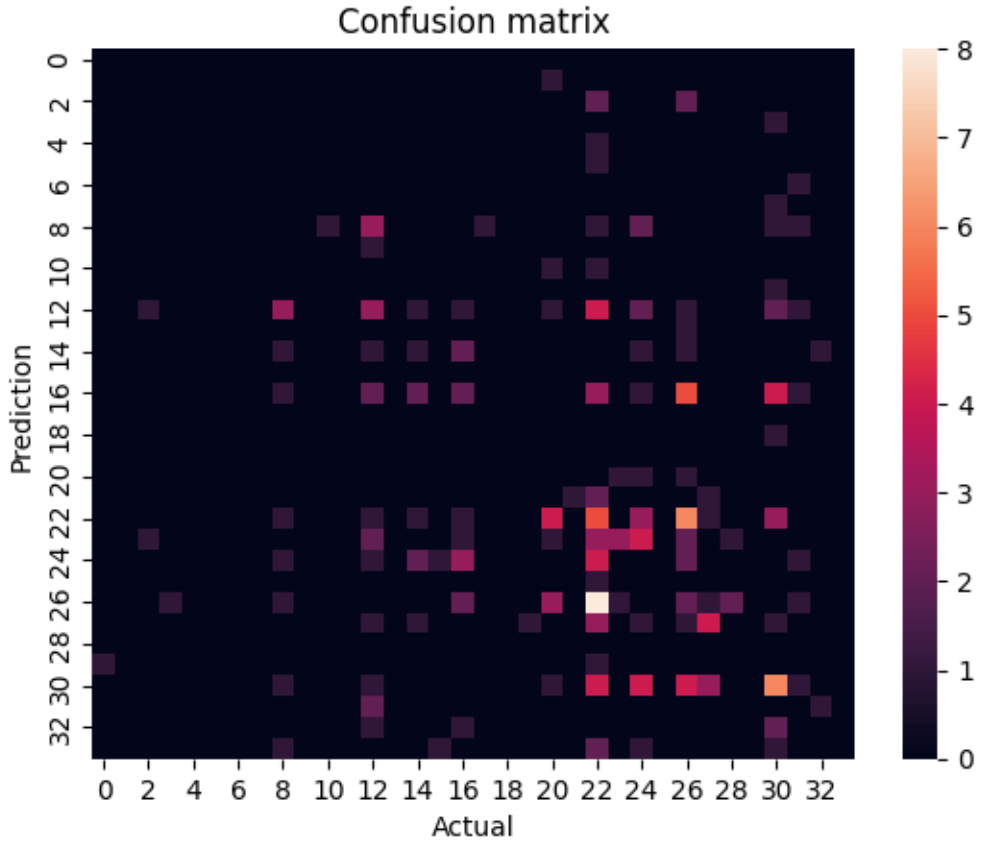
$$F_1 = \frac{1}{n}\sum\frac{2P_XR_X}{P_X+R_X} \quad (12)$$

Missä n on luokkien lukumäärä, P_X on yksittäisen luokan precision eli sisäinen tarkkuus ja R_X on yksittäisen luokan recall eli herkkyys, jotka sitten summataan yhteen kaavan 12. mukaisesti. F1-arvojen keskiarvo lasketaan kaavalla, missä harmoninen keskiarvo lasketaan aritmeettisista keskiarvoista. [27].

$$\mathbb{F}_1 = 2\frac{\left(\frac{1}{n}\sum P_X\right)\left(\frac{1}{n}\sum R_X\right)}{\frac{1}{n}\sum P_X + \frac{1}{n}\sum R_X} \qquad (13)$$

Missä n:ät ja P_X ja R_X tarkoittavat samoja asioita kuin keskiarvoistetussa F1-arvossa kaavassa 12.

Sekaannusmatriisin avulla on mahdollista esittää mallin ennustukset luokittain. Kuvassa 4. on esimerkki satunnaismetsä tyyppisen luokittelijan sekaannusmatriisista. On mahdollista havaita, mitä luokkia malli on ennustanut kullekin luokalle. Kuvassa luokat ovat koodattu todellisista luokista numeeriseen muotoon.



Kuva 4. Esimerkkikuva satunnaismetsän ennustuksen sekaannus matriisista

4.5 Ajoympäristö

Opinnäytetyön kehitysvaiheessa algoritmin kehittämis- ja ajoympäristönä toimii Jupyter Notebook. Jupyter on voittoa tavoittelematon, avoimen lähdekoodin projekti, joka on syntynyt 2014 IPython-projektin pohjalta. Jupyter notebook-dokumenteilla eli muistioilla on mahdollista ajaa in-

teraktiivista ohjelmointiympäristöä, jotka soveltuvat tieteelliseen laskentaan ja muuhun datatieteelliseen kehittämistyöhön. Jupyteria kehitetään avoimesti GitHub-nimisellä verkkosivustolla. [28]. Hankkeen myöhemmässä vaiheessa olisi mielekästä, että luistovoiteen suosituksia pystyisi tekemään myös mobiilisovelluksesta.

5 Luistovoiteen suosittelualgoritmi

Tässä opinnäytetyön luvussa esitellään aineisto, algoritmin tavoite, algoritmien rakenteet ja algoritmien tulokset. Aineiston esittely luvussa esitellään, mitä muuttujia aineistossa on, mitenkä ne jakautuvat ja onko niissä puuttuvia arvoja. Aineiston esittely luvuissa esitellään myös se, minkälainen esikäsittely aineistolle tehdään algoritmeja varten sekä myös se minkälaisena aineisto syötetään algoritmeille. Algoritmin tavoite luvussa esitellään se, mitenkä algoritmeja arvioidaan ja mikä niiden tavoite on. Algoritmin rankenne -luvussa esitellään opinnäytetyössä kehitettyjen algoritmien rakennetta. Algoritmien tulokset luvussa esitellään algoritmien tuloksia.

5.1 Aineiston esittely

Suosittelualgoritmin aineisto muodostuu testitietokannasta ja testipaikkojen tiedoista. Testitietokanta pitää sisällään tiedon testin päivämäärästä, paikasta, ilmanlämpötilasta, lumen lämpötilasta 1 cm:n syvyydessä, tiedon käytetystä suksesta, tiedon käytetyistä voiteista, tiedon käytetyistä käsikuvioista, testintekijän nimen, tiedon testin tyypistä, tiedon konekuviosta, tiedon suhteellisesta ilmankosteudesta, tiedon kastepistelämpötilasta, tiedon sähkönjohtamisesta lumenpinnassa, joka on nimetty uramuuttujaksi aineistossa ja tiedon testiin lisätystä lisäselitteestä. Testipaikkojen aineisto pitää sisällään tiedon testipaikasta, maasta, koordinaatista (GPS) ja korkeuden merenpinnasta. Testipaikkojen tiedot voidaan yhdistää testitietokannan aineistoon käyttäen testin paikkatietoa ja testipaikkojen paikkatietoa. Aineistossa näyterivejä on hieman alle 9 000.

Yhteen testitietokannan testiin voi kuulua useampi määrä suksipareja 1–12 väliltä, mutta yleisimmät yhdessä testissä olleet suksipari määrät ovat 4 ja 8. Uniikkeja suksia testiaineistossa on 366 kpl, jotka kaikki ovat vapaan hiihtotavan suksia. Uniikkeja konekuvioita aineistossa on hieman alle 180 kappaletta. Erilaisia voiteenlaittotapa ja -seos -yhdistelmiä aineistosta löytyy yli 1400 kappaletta ja uniikkeja käsikuviointeja hieman alle 550 kappaletta. Testejä on suoritettu noin 40:ssä eri paikassa, joten korkeus merenpinnasta testien välillä vaihtelee 10 metristä 2028 metriin. Ilman lämpötila testattujen testien välillä vaihtelee -22.5 celsiusasteen ja +18 celsiusasteen välillä. Suhteellinen ilmankosteus 0.13 ja 0.99 välillä. Lumen lämpötila 1 cm:n syvyydessä -22 ja -0.1 celsiusasteen välillä. Kastepiste -30.26 ja +5.63 celsiusasteen välillä. Sähkönjohtavuus (kuvaa veteen liuenneiden suolojen määrää [29.]) lumen pinnassa vaihtelee 12 ja 99 välillä. Testin tyyppinä tes-

tiaineistossa on joko voidetesti, kuviotesti tai kalibrointi. Testin tuloksessa on mukana myös tunneominaisuus 1–5 asteikolla. Nämä voivat toimeksiantajan mukaan olla tärkeitä arvoja, jos arvona on 1 tai 2, koska tällöin tunne sukseen on subjektiivisesti ollut hyvä. Lisäksi testeistä löytyy ominaisuus nimeltä Ero% kärkeen, joka on vain hyödyllinen toimeksiantajan mukaan verrattaessa sitä testipakkojen sisällä. Voiteiden ja kuvioiden tiedoissa on myös tieto siitä, mihin alueihin voidetta on levitetty ja mitenkä. Erilaisia testiaineistossa käytettyjä voiteen laittotapoja on mm. polttaminen, raudoitus, liotus, fliisaus, käsikorkki ja konekorkki. Uniikkeja ensimmäisenä voiteena käytettyjä voiteita on noin 600, mutta näistä n. 130 voidetta on käytetty vain kerran ensimmäisenä voiteena. Noin 3700 testinaineiston näytteessä eli melkein puolessa ei ole aineistoon merkattu ollenkaan käytettyä voidetta. Toimeksiantajan mukaan näissä testeissä on kuitenkin käytetty voidetta ja se on ollut kaikissa suksissa sama. Voiteita käytetyissä testeissä puolet testeistä on tehty 50 käytetyimmällä voiteella. Hieman yli 6100 näyterivillä ei ole aineiston mukaan käytetty ollenkaan käsikuviointia. Tarkkaa määrää testiaineistossa käytetyistä voiteista on vaikea sanoa, koska sama voide on voitu kirjoittaa monella eri tavalla tai väärin. Sama pätee myös käsi- ja konekuviointeihin ja jopa paikkojen tietoihin. Voidetesteissä oli käytetty 218 eri suksea, joista 41 suksea oli käytetty yli puolista testeistä.

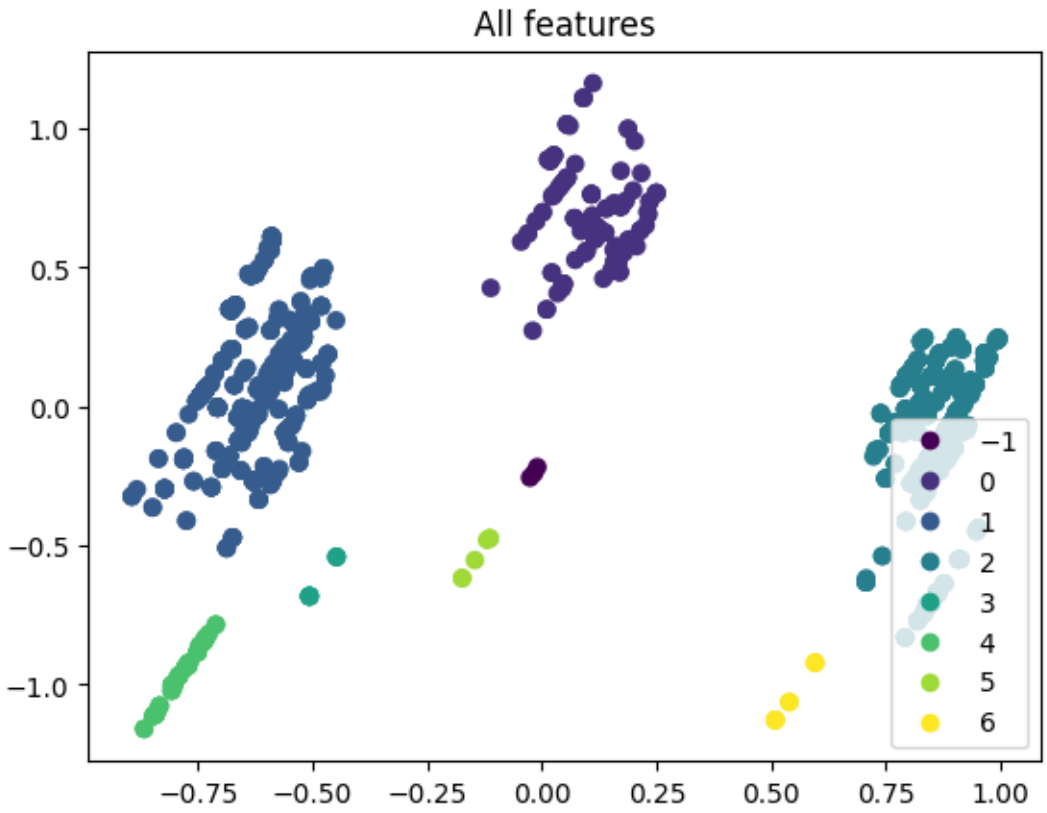
Taulukossa 1. esitellään koko lähdeaineisto taulukkomuodossa. Taulukossa esitellään muuttujan nimi, tietotyyppi, numerollisille muuttujille minimi- ja maksimiarvo, ei-numeerisille muuttujille uniikkien arvojen määrä ja kaikille muuttujille aineistosta puuttuvien rivien määrä.

Taulukko 1. Lähdeaineiston esittely

Muuttujan nimi	**Tietotyyppi**	**Minimiarvo**	**Maksi-miarvo**	**Uniikit arvot**	**Puuttuvat rivit**
Päivämäärä	Merkkijono	30.10.2019	16.4.2023	354	0
Paikka	Merkkijono	-	-	39	0
Ilmanlämpötila	Liukuluku	-22.5	+18	-	0
Lumen lämpötila 1 cm	Liukuluku	-22	-0.1	-	0
Suksen tunniste	Merkkijono	-	-	366	0
Voiteet	Merkkijono	-	-	1450	3695
Käsikuviointi	Merkkijono	-	-	661	6120
Testintekijä	Merkkijono	-	-	5	0
Testin tyyppi	Merkkijono	-	-	3	0
Konekuviointi	Merkkijono	-	-	178	6109
Suhteellinen ilman-kosteus	Liukuluku	0.13	0.99	-	0
Kastepisteen läm-pötila	Liukuluku	-30.26	5.63	-	0
Sähkön johtaminen lumen pinnassa (Ura)	Kokonaisluku	12	99	-	1825
Lisäselite	Merkkijono	-	-	58	116
Maa	Merkkijono	-	-	29	0

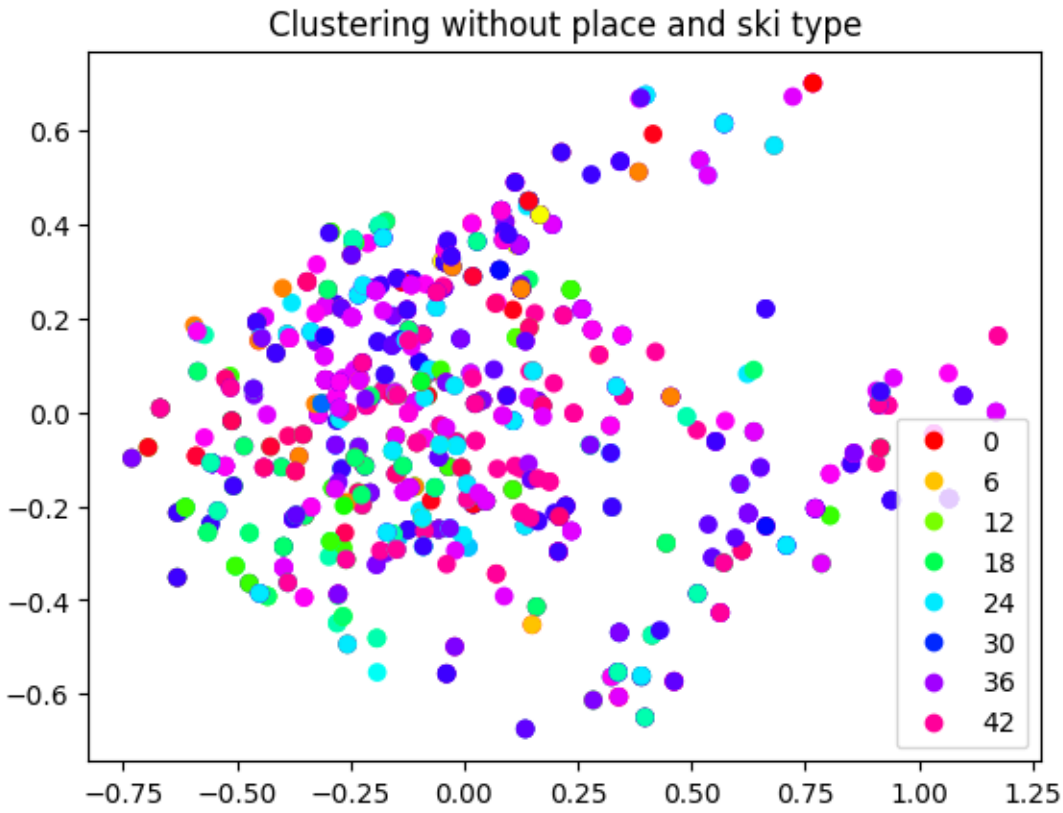
Koordinaatit	Merkkijono (GPS)	-	-	39	0
Korkeus merenpinnasta	Kokonaisluku	10	2028	-	0
Tunne	Kokonaisluku	1	5	-	5843
Ero kärkeen	Liukuluku	0	2.89	-	6
Sijoitus testissä	Merkkijono	-	-	67	2863

Kuvassa 5. on pistekaavio klusteroidulle aineistolle. Siinä on siis kokeiltu ryhmitellä aineistoa pääkomponenttianalyysin (eng. principal components analysis, PCA) ja klusterointialgoritmin avulla, missä ryhmittymät ovat muodostuneet suksen nimessä olevien kirjain yhdistelmien (MH, AH ja YH) ja muutaman testipaikan mukaan. Klusterointi on suoritettu DBSCAN (Density-based spatial clustering of applications with noise) klusterointialgoritmin avulla. Se tehdään niin, että pääkomponenttianalyysillä valitaan aineistosta ne ominaisuudet, joille projisoituna data tuottaa suurimman varianssin [30]. Sen jälkeen ne on klusteroitu DBSCAN klusterointi-algoritmilla. Klusterointialgoritmin on löytänyt aineistosta seitsemän eri ryhmää ja muutaman pisteen, jotka eivät kuulu oikein mihinkään ryhmään, kun DBSCAN klusterointialgoritmin epsilon parametri on asetettu 0.20, joka määrittää kahden pisteen maksimimatkan, jotta ne luokitellaan samaan ryhmään.



Kuva 5. Aineiston klusterointi kaikilla muuttujilla

Kuvassa 6. on tehty samat prosessit aineistolle kuin kuvassa 5. sillä poikkeuksella, että aineistosta on poistettu muuttujat, jotka kuvaavat suksen tunnisteita ja testipaikkoja. Lisäksi pistekaavio on värjätty voidemerkkien mukaan. Klusteroinnin avulla pyrittiin etsimään samantyylisiä voiteita ja ryhmittämään niitä. Kuvaan tehdystä pääkomponenttianalyysistä suurimman varianssin pääkomponentteihin tuottavat muuttujat lumen lämpötilasta 1 cm syvyydessä ja korkeus merenpinnasta.



Kuva 6. Aineiston klusterointi ilman paikka- ja suksen tunniste -muuttujia

5.2 Esikäsittely algoritmia varten

Datan esikäsittely algoritmia varten alkaa väärinkirjoitettujen testipaikkojen korjaamisesta. Osa testipaikoista on kirjoitettu osittain väärin, kokonaan väärin tai lyhennetty lyhyempään muotoon. Tällaiset korjattavat tiedot pitää ensin löytää aineistosta ja sitten korjata. Onneksi testiaineiston lisäksi algoritmin kehityksessä on mukana myös aineisto testipaikoista, jonka mukaan pystyy havaitsemaan, mikä on testipaikkojen oikea nimi. Testiaineistosta on korjattu paikkojen nimiä 31 kappaletta. Toimeksiantajan uuteen testiaineiston keräysalustaan testipaikat kannattaa olla käyttäjälle kirjoitettuna valmiiksi, jotta vältytään paremmin mahdollisilta kirjoitusvirheiltä. Uuden testipaikan tullessa uuteen järjestelmään, pääkäyttäjä tai vastaavaa muokkaus roolia hallitseva kävisi lisäämässä sen sitten listaan testipaikoista.

Aineistossa selkeästi poikkeavia arvoja on vain yhdessä muuttujassa, joka kuvaa lumen kosteutta (Ura-muuttuja). Mutta tämän muuttujan arvoja ei ole välttämättä mielekästä ruveta rajaamaan,

koska toimeksiantajan mukaan uran kosteus on yksi merkittävimmin luistoon vaikuttavista muuttujista. Muista numeerisista muuttujista hieman kvartiiliväli -menetelmällä poikkeavia arvoja on muuttujissa ilman lämpötila, suhteellinen ilmankosteus, lumen lämpötila 1 cm:n syvyydessä ja kastepiste. Mutta nämä arvot poikkeavat niin vähän, että niiden rajaamisesta ei saataisi erityisempää hyötyä aineiston valmistelussa algoritmia varten. Kvartiiliväli (IQR) on yläkvartiilin (Q_3) ja alakvartaalin (Q_1) erotus, josta lasketaan ylempi viiksi lisäämällä yläkvartiiliin 1.5 kertaa kvartiiliväli ja alempi viiksi vähentämällä alakvartiiliista 1.5 kertaa kvartiiliväli. Tilastollisesti poikkeavia arvoja ovat näiden viiksien ulkopuolelle jäävät arvot. Visuaalisesti tämän poikkeavien arvojen suodattamistavan voi havaita käyttämällä laatikkokuvaajaa (eng. box plot), jonka tunnetuksi esitteli John Tukey kirjassaan Exploratory Data Analysis vuonna 1977 [31]. Kvartiiliväli menetelmässä käytetään kertoimena 1.5, koska normaalisti jakautuneessa satunnaismuuttujassa se sisältäisi 99.72 % havainnoista, mutta rajaisi kaiken muun pois [32].

Ennen kuin aineistossa käytetyt muuttujien arvot voidaan syöttää algoritmille ne pitää skaalata. Muuttujien skaalaaminen pitää tehdä siksi, että muuten algoritmi painottaisi liikaa suurempia arvoja. Muuttujan skaalaamiseen on useita menetelmiä. Valittuun skaalaamismenetelmään vaikuttaa lähinnä se, mihin muotoon muuttuja halutaan skaalata. Yleisimmät menetelmät ovat absoluuttinen maksimiskaalaus, minimi-maksimi-skaalaus, normalisointi, standardointi ja vankka skaalaus. Menetelmästä huolimatta skaalauksen päämäärä on saada muuttuja skaalattua tietylle välille, esimerkiksi 0 ja 1 välille [33].

Muuttujan minimi-maksimi-skaalaus (eng. Min-Max scaling) onnistuu kaavalla [33],

$$X_{scaled} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (14)$$

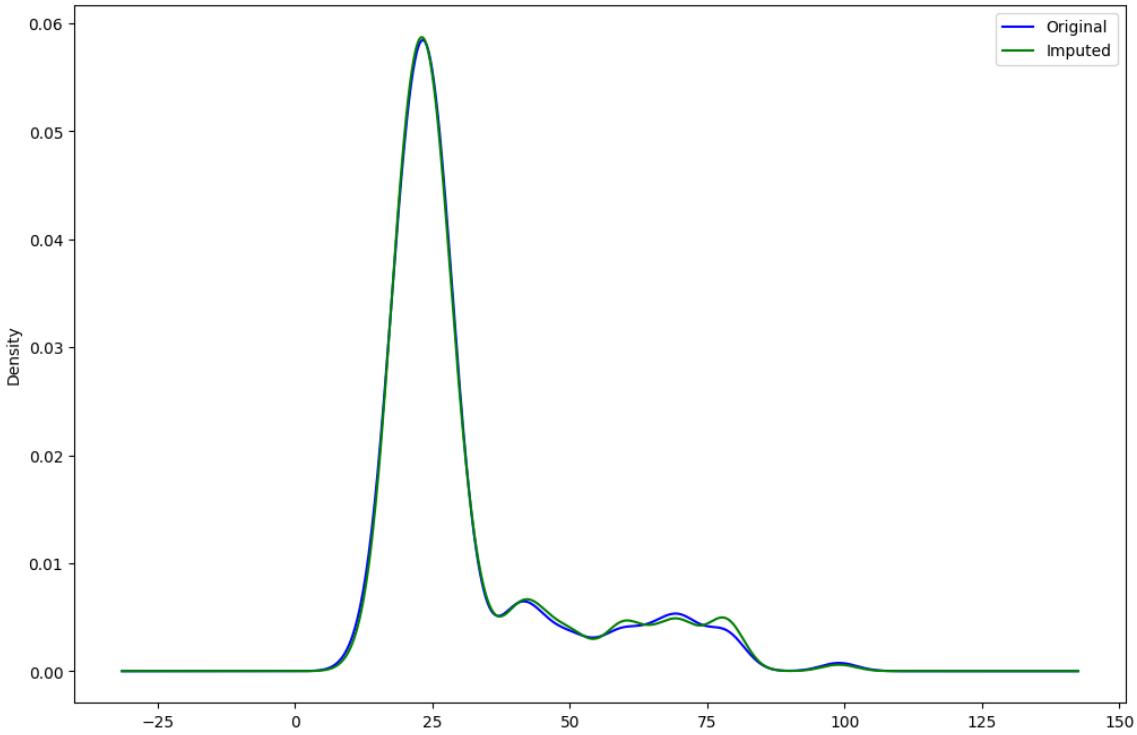
Missä X_i on käsiteltävä arvo, X_{min} on muuttujan pienin arvo ja X_{max} on muuttujan suurin arvo.

Algoritmia varten käytetyissä muuttujista skaalataan minimi-maksimi-skaalauksella muuttujat ilman lämpötila, suhteellinen ilmankosteus, lumen lämpötila 1 senttimetrissä, kastepiste ja ura. Lisäksi ero kärkeen -muuttuja skaalataan testipakkojen sisällä niin, että testipakat, jossa on enemmän kuin 5 suksea suurimman eron saanut suksi saa suhteutetuksi eroksi 1 ja pienimmän eron saanut suhteutetuksi eroksi 0. Ja niissä testipakoissa, joissa on vähemmän kuin 5 suksea suurimman eron saanut suksi saa suhteutetuksi eroksi 0.5 ja pienimmän eron saanut suhteutetuksi eroksi 0.

Tunne -muuttujan arvot ovat jo järjestyksellisessä muodossa niin, että pienemmät arvot ovat parempia. Ongelmana on vain se, että tunne -muuttuja pitää sisällään paljon puuttuvia arvoja, jotka

pitää korvata, että muuttujan käyttäminen olisi hyödyllistä. Koska tunne -muuttujan ja suhteutettu ero -muuttujan korrelaatio on vahva, noin 0.75 pearsonin korrelaatiokertoimella, voidaan tunne -muuttujan puuttuvia arvoja korvata suhteutettu ero muuttujan avulla ja toisin päin. Vahva positiivinen korrelaatio 0.70–0.89 kahden muuttujan välillä tarkoittaa sitä, että kun toisen muuttujan arvot kasvavat toisenkin muuttujan arvot. [34]. Koska tunne -muuttujan arvot ovat asteikolla 1-5 ja suhteutettu ero -muuttujan arvot ovat asteikoilla 0–1. Voimme puuttuvat tunne -muuttujan arvot asettaa yhdeksi, kun suhteutettu ero on 0.00–0.19, kahdeksi silloin kuin suhteutettu ero on 0.20–0.39, kolmeksi silloin, kun suhteutettu ero on 0.40–0.59, neljäksi silloin, kun suhteutettu ero on 0.60–0.79 ja viideksi silloin, kun suhteutettu ero on 0.80–1.00. Samaa menetelmää voimme käyttää myös toisin päin, jos testissä on merkattu tunne, mutta ei suhteutettu ero -muuttujaa.

Tunne ja suhteutettu ero -muuttujien lisäksi puuttuvia arvoja löytyy myös ura -muuttujasta. Niin kuin aikaisemmilla muuttujilla oli korrelaatiota keskenään, niin myös ura -muuttujalla on korrelaatioita muihin aineistossa löytyviin muuttujiin. Ura -muuttujalla on noin 0.5:n korrelaatio muuttujiin lumen lämpötila 1 cm syvyydessä, kastepiste ja ilman lämpötila. Tämän takia puuttuvat ura -muuttujan arvot olisivat siis mielekästä paikata hyödyntäen muita muuttujia, koska muuttujien välillä on hieman korrelaatiota. Tämä on mahdollista tehdä imputoinnnin avulla tarkemmin ottaen MICEn eli Multiple Imputation by Chained Equations avulla. Suomennettuna se on moninkertainen imputointi ketjutetuilla yhtälöillä. MICE-imputointi tapahtuu niin, että ensin puuttuvat arvot korvataan esimerkiksi muuttujan keskiarvolla. Tämän jälkeen regressiomalliin syötetään muuttuja, josta puuttuu arvoja riippuvaisena muuttujana ja muuttujan kanssa korreloivat muuttujat riippumattomina muuttujina regressiomalliin. Tämän jälkeen muuttujan puuttuvia arvoja ennustetaan regressiomallin avulla ja paikataan näillä ennustetuilla arvoilla. [35]. Python-ohjelmointikielelle on olemassa Miceforest-kirjasto, jonka ImputationKernel-luokan avulla on mahdollista tehdä imputointi. Liitteessä 1. on lähdekoodi ura muuttujan imputoinnista. Kuvassa 7. on kuvaaja alkuperäisen ja imputoidun ura -muuttujan arvojen jakaumasta.



Kuva 7. Ura muuttujan alkuperäisestä ja imputoidusta jakautumisesta

Puuttuvien arvojen korvaamisen onnistumista voidaan tutkia katsomalla, miten imputoidun ja alkuperäisen muuttujan jakautuminen eroaa. Kuvassa 7. sinisellä viivalla muuttujan alkuperäinen jakautuminen ja vihreällä imputoitu muuttuja. Imputoidun ja alkuperäisen muuttujan jakautumiset menevät lähes päällekkäin. Muuttujan imputointi on siis onnistunut.

Koska koneoppimismallit edellyttävät, että mallille syötetyt arvot ovat numeerisia, ei-numeeriset arvot pitää muuttaa numeeriseen muotoon [36]. Ei-numeerinen eli kategorinen tieto voidaan jakaa kahteen eri luokkaan, järjestykselliseen ja nimelliseen tietoon. Järjestyksellisessä tiedossa tiedetään siis luokkien järjestys. Esimerkiksi tieto henkilön korkeimmasta suoritetusta tutkinnosta on järjestyksellinen kategorinen tieto. Nimellistä kategoriatietoa taas ei pysty laittamaan järjestykseen. Kategorisentiedon muuttamiseksi numeeriseen muotoon on olemassa muutamia menetelmiä. Yleisin menetelmä kategorisentiedon muuttamisessa numeeriseen muotoon on One hot -koodaus (eng. One-Hot Encoding, OHE). Siinä kategorisen muuttujan kaikki uniikit näytteet koodataan omaksi binääriseksi sarakkeekseen. One hot -koodauksen yhtenä ongelmana on se, että sarakkeiden määrä koodatussa aineistossa saattaa kasvaa merkittävästi, jos kategorinen muuttuja pitää sisällään paljon uniikkeja näytteitä. Toinen menetelmä kategorisentiedon koodaamiseksi on nimiö -koodaus (eng. label encoding), siinä kategorisen muuttujan uniikit näytteet koodataan numeroksi. Haittapuolena tässä on se, että koneoppimismalli saattaa oppia painottamaan näytteitä koodatun numeron mukaan [37].

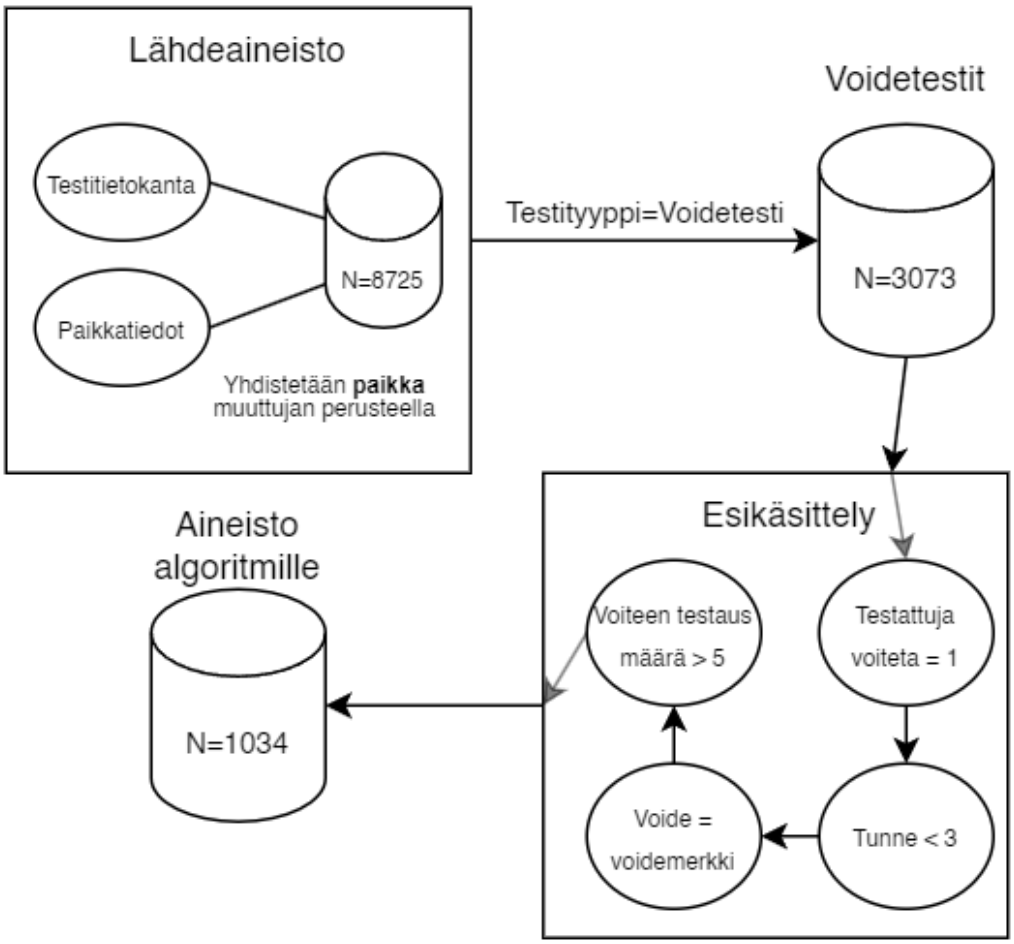
5.3 Aineisto algoritmille

Aineiston lopullinen muoto algoritmin opetukseen on sellainen, missä muuttujiksi on valittu ilman lämpötila, suhteellinen ilmankosteus, lumen lämpötila 1 cm:n syvyydessä, kastepiste, sähkönjohtaminen lumenpinnassa, testipaikan korkeus merenpinnasta, suksen tunniste maastohiihdon, ampumahiihdon ja yhdistetyn hiihdon väliltä ja testipaikka. Ennustettavana muuttujana on testissä käytetty voide, joka on yksinkertaistettu vain voidemerkkiin. Kussakin voidemerkissä voiteiden määrä vaihtelee 4 ja 94 välillä. Osassa voidemerkeissä erot voiteiden välillä voivat olla eri versio samannimisestä voiteesta tai sama voide on vain kirjoitettu eri tavalla, mikä nostaa uniikkien voiteiden määrää. Tämän takia kohdemuuttujana käytetään vain voidemerkkiä. Ennustustehtävän helpottamiseksi opetus- ja testiaineistoon on valittu vain sellaiset näyterivit, jossa testin tunne -muuttujan arvo on ollut 1 tai 2 ja testeistä on valittu vain sellaiset näytteet, missä on testattu vain yhtä voidetta. Tämän takia aineisto muodostuu voidenäytteistä, jotka ovat omissa testeissään hyväksi todettuja voiteita kyseisille keliolosuhteille. Lisäksi testeistä on valittu vain sellaiset näyterivit, missä voidemerkkiä on testattu useammin kuin viidesti. Opetus- ja testiaineiston jako on tehty Scikit-learn-kirjaston train_test_split-funktiolla, jonka avulla esikäsitellystä aineistosta 80 % on jaettu mallin opetukseen ja 20 % mallin testaukseen satunnaisesti. Opetus- ja testiaineiston näytteiden lopullinen yhteenlaskettu määrä algoritmia varten on noin 1000 näyteriviä. Aineistosta ennustettavien voidemerkkien näytteiden määrä vaihtelee 135 ja 7 välillä, niin että kolmelle eniten aineistossa esiintyvälle voidemerkille on yli 100 näytettä ja kolmelle vähiten aineistossa löytyvälle voidemerkille alle 10 näytettä. Lopulle 12 ennustettavalle voidemerkille aineistossa on keskimäärin 53 näytettä. Taulukossa 2. näkyy algoritmille valittujen muuttujien nimi, tietotyyppi, jakauma ja esikäsittely -menetelmä.

Taulukko 2. Aineiston esittely

Nimi	**Tietotyyppi**	**Jakauma**	**Esikäsittely**
Ilman lämpötila	Liukuluku	-22.5 - 14	MinMax scaling
Suhteellinen kosteus	Liukuluku	0.13 – 0.99	MinMax scaling
Lumen lämpötila 1 cm syvyydessä	Liukuluku	-21 - -0.1	MinMax scaling
Kastepiste	Liukuluku	-29.19 – 5.63	MinMax scaling
Ura (Sähkön johtaminen lumessa)	Liukuluku	13 – 99	MinMax scaling
Korkeus	Kokonaisluku	10 – 2028	MinMax scaling
Suksen tunniste	Merkkijono	AH, YH, MH	OHE
Paikka	Merkkijono	34 kpl	OHE
Voidemerkki (Kohdemuuttuja)	Merkkijono	18 kpl	Label encoding

Kuvassa 8. visualisoituna dataputki lähdeaineistosta algoritmille syötettävään aineistoon. Mistä voimme havaita aineiston määrän vähentyvän, kun lähdeaineistoa tehdään algoritmille hyödyllistä aineistoa.



Kuva 8. Aineiston dataputki

5.4 Tavoite

Tässä opinnäytetyön alaluvussa esitellään sitä, mikä koneoppimisalgoritmien tavoite on ja sitä mitenkä tähän tavoitteeseen pääsyä arvioidaan.

5.4.1 Algoritmin arviointi

Ennustusalgoritmeja arvioidaan ja vertaillaan luvussa 4.4 esiteltyjen arviointimetriikoiden avulla. Niitä ovat makrokeskiarvoistetut sisäinen tarkkuus, herkkyys ja kaavan 12. mukainen makrokeskiarvoistettu F1-arvo sekä mallin tarkkuus. Lisäksi malleille arvioidaan ennustuskykyä testeissä eniten esiintyvien voiteiden ennustamiseen, jotka ovat REX, STAR ja TOKO.

5.4.2 Algoritmin tavoite

Algoritmin tavoitteena on luistovoiteen suosittelu keliolosuhteisiin perustuen. Eli keliolosuhteet ja muut testiin liittyvät muuttujat syöttämällä algoritmiin olisi sille mahdollista saada sopiva luistovoide, joka mahdollisesti myös toimisi mahdollisimman hyvin. Jotta algoritmi toimii paremmin, kuin satunnaisesti voiteita arvaamalla, täytyy algoritmin saada parempi tulos kuin vertailuna käytettävä Scikit-learn Python-kirjaston DummyClassifier-luokittelija saa tulokseksi käyttäessään luokittelu strategiana most_frequent eli eniten toistuvaa. Silloin luokittelija ennustaa testiaineiston luokan sen mukaan, mikä on ollut opetusaineistossa eniten toistuva luokka.

DummyClassifier sai ennustus tehtävästä tarkkuudeksi 13 %, makrokeskiarvoistetulle sisäiselle tarkkuudelle 0.7 %, makrokeskiarvoistetuksi herkkyydeksi 5.5 % ja makrokeskiarvoistetuksi F1-arvoksi 1.3 %. Koska luokittelijan ennustusstrategiana on ennustaa suosituinta testeissä käytettyä voidetta, muodostui ennustuksen tarkkuus vain REX-voidemerkin ennustamisesta. Voidemerkki sai sisäiseksi tarkkuudeksi 13 %, herkkyydeksi 100 % ja F1-arvoksi 0.23.

5.5 Algoritmin rakenne

Liitteessä 2. on lähdekoodit koneoppimismalleille

5.5.1 Satunnaismetsä

Satunnaismetsä luokittelija koneoppimismenetelmälle on mahdollista asettaa Pythonin Scikit-learn-kirjastossa parametreiksi puiden lukumäärä (n_estimators), funktio mittaamaan halkaisujen laadukkuutta (criterion), maksimisyvyys (max_depth), maksimi -muuttujien lukumäärä

(max_features), minimiepäpuhtaus lasku (min_impurity_decrease), luokan painokerroin (class_weight), satunnaistila (random state) ja muita vähemmän ennustukseen vaikuttavia parametreja tai toisenlaiseen ennustus tehtäviin tarkoitettuja parametreja.

Opinnäytetyössä käytettyyn satunaismetsä-koneoppimismalliin parametreiksi on asetettu satunnaismetsän satunnaiseksi tilaksi 42, jotta tulokset ovat uudestaan toistettavia. Muut parametrit ovat Scikit-learn-kirjaston oletusparametrien arvojen mukaisia, milloinka esimerkiksi halkaisujen laadukkuutta kertova funktio on asetettu giniksi.

5.5.2 Neuroverkko

Algoritmin kehittämisessä käytetyt neuroverkko mallit ovat Scikit-learn-kirjaston MLP (Multilayer Perceptron) eli monikerroksinen perseptroniverkko luokittelumalli ja Googlen kehittämällä TensorFlow-kirjastolla kehitetty oma neuroverkkorakenne.

Scikit-learn-kirjaston MLP-luokittelijassa säädettäviä parametreja on muun muassa piilokerrosten koko (eng. hidden layer sizes), piilokerroksen aktivointifunktio (eng. activation function), perseptronien painojen optimointimenetelmä (eng. solver), oppimisnopeus (eng. Learning rate), neuroverkolle opetettavan erän koko (eng. Batch size), L2-rekulalisaatiotermin vahvuus (alpha), iteraatioiden määrä (eng. Max-iter) sekä muita säädettäviä parametreja, jotka on muutettavissa muiden asetettujen parametrien perusteella. Näistä ehkä tärkeimpänä parametrina on perseptronien eli neuronien painojen optimointimenetelmä. Vaihtoehtoja MLP-mallissa tällä parametrille on LBFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algoritmi, joka perustuu quasi-Newton menetelmiin, SDG (stochastic gradient descent) stokastinen gradienttilasku ja Adam-optimoija (Adaptive Moment Estimation), joka perustuu stokastiseen gradienttiin [38].

Opinnäytetyössä käytetylle MLP-luokittelijalle asetettiin parametreiksi maksimi-iteraatioksi 10 000 ja satunnaiseksi tilaksi 42, jotta tulokset ovat uudestaan toistettavissa. Muut parametrit on asetettu Scikit-learn-kirjaston oletus parametri arvojen mukaisesti. Jolloin esimerkiksi optimointi -menetelmäksi tulee Adam-optimoija ja aktivointi funktioksi ReLU (Rectified Linear Unit) sekä piilokerrosten kooksi 100.

TensorFlow-kirjastolla Keras-rajapinnalla kehitettyyn neuroverkkomalliin on mahdollista määritellä neuroverkon kerrokset (eng. layers), kerroksille neuronien lukumäärät, kerroksille aktivaatiofunktiot, mallin optimointi -menetelmä sekä monia muita eri parametreja ja rakenteita.

Opinnäytetyössä neuroverkoksi TensorFlow-kirjastolla rakennettiin 4-kerroksinen neuroverkko. Neuroverkossa ensimmäinen kerros on opetusaineistossa muuttujien määrän mukainen kerros neuroneja. Toinen kerros on täysin kytketty (eng. dense) 16-neuroninen kerros, missä aktivointifunktiona on ReLU. Täysin kytketty kerros tarkoitta sitä, että kerroksen kaikki neuronit ovat kytkettyinä kaikkiin edellisen kerroksen neuroneihin [39] ja ReLU-aktivointifunktio on funktio, joka palauttaa neuronin arvon, jos arvo on positiivinen muuten nollan [40]. Kolmantena kerroksena on pudotuskerros (eng. dropout layer), missä pudotetaan satunnaisesti 50 % kerroksen neuroneista. Tämä estää mallin ylisovittumista ja parantaa mallin generalisointia [41]. Viimeisenä ulostulokerroksena on kerros, missä neuronien lukumäärä on sama kuin ennustettavien voidemerkkien lukumäärä on aineistossa. Kerroksen aktivointifunktiona on softmax-funktio. Softmax-funktio lasketaan kaavalla 15. [42].

$$\sigma(x) = \frac{e^x}{\sum e^x} \qquad (15)$$

Missä e on Eulerin numero ja x on vektori sisään tulevista arvoista.

5.6 Algoritmin tulokset

Tässä opinnäytetyön alaluvussa esitellään koneoppimismallien tuloksia.

5.6.1 Satunnaismetsä

Satunnaismetsä-luokittelukoneoppimismalli osasi luokitella testiaineiston 207 näytteestä 30 oikein, silloin kun malli ennusti luistovoiteen voidemerkkiä. Prosentuaalisesti tämä on 14.5 % oikein ennustettu. Scikit-Learn Python-kirjaston satunnaismetsän koneoppimismallista on myös mahdollista saada opetusaineiston muuttujien epäpuhtauspohjainen tärkeys ulos hakemalla feature_importances_ attribuuttia opetetusta mallista. Epäpuhtauspohjainen muuttujan tärkeys lasketaan kunkin satunnaismetsän puun sisältämän keskiarvon ja keskihajonnan kertymän vähenemisestä kussakin puussa. [43.] Opetetun mallin tärkeimmät muuttujat voiteen ennustamiseen olivat kastepiste, ilmanlämpötila, sähkön johtavuus, suhteellinen ilmankosteus ja lumen lämpötila 1 cm:n syvyydessä. Nämä muuttujat selittivät kukin noin 15 % ennustuksesta eli yhteensä 75 %. Seuraavaksi tärkeimpänä muuttujana oli korkeus, joka selitti noin 6 % ennustuksesta. 2–1.5 %

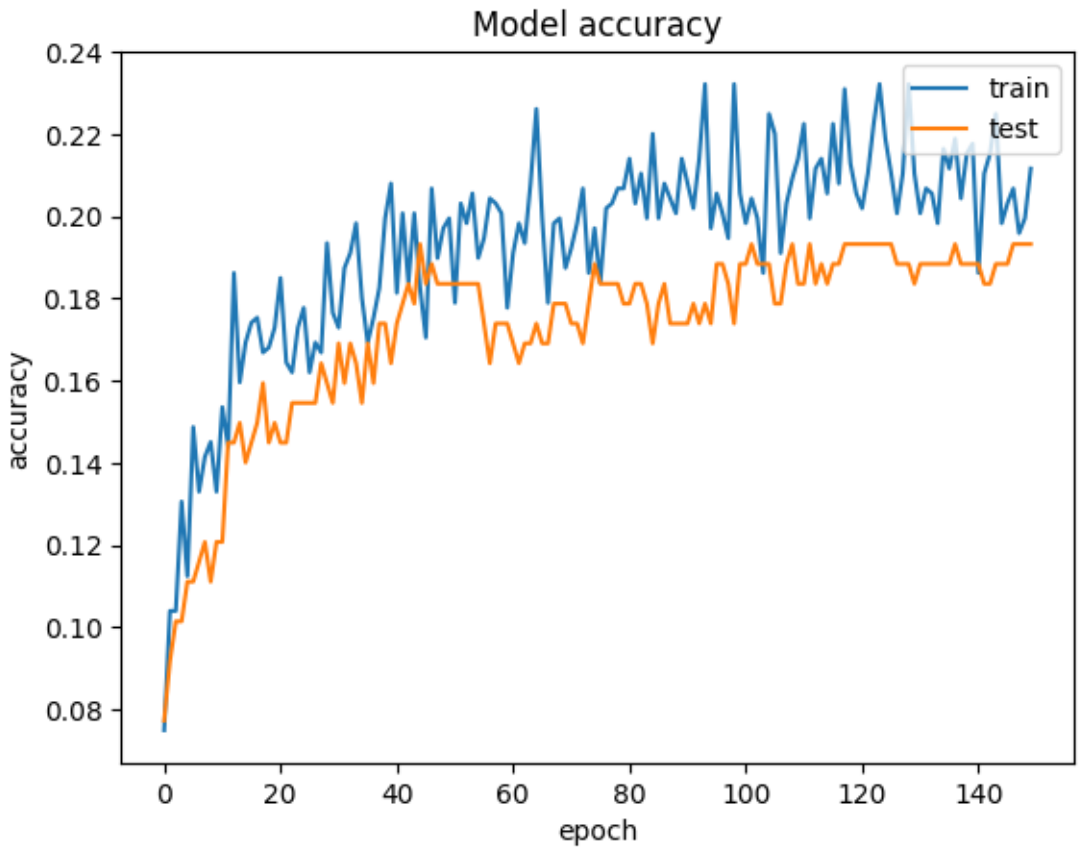
selittävyydet ennustukseen tulivat testissä käytetyn suksen tunnisteesta. Loput voiteen ennustuksen selittävyydestä tulivat paikkamuuttujan perusteella. Tosin epäpuhtauspohjainen muuttujan tärkeys voi olla harhaan johtava muuttujan selittämiskyvylle, jos se sisältää paljon uniikkeja arvoja. Epäpuhtauspohjainen muuttujan tärkeys lasketaan myös opetusaineiston pohjalta, mikä saattaa vaikuttaa muuttujien tärkeyteen. Muuttujien tärkeyden voi myös laskea muuttujien permutaation pohjalta, jonka on tarkoitus päihittää epäpuhtauspohjaisen muuttujan tärkeyden laskemistavan rajoitteet. [44.] Permutaatiopohjaisen muuttujien tärkeyden selittävimpinä muuttujina on ampumahiihto ja maastohiihto nimiset suksitunnisteet, lumen lämpötila 1 cm:n syvyydessä ja Vuokatti, Beijing, Lillehammer ja Dresden toimiessa testipaikkana. Mallin tulosta huonontavana muuttujina permutaationpohjainen muuttujatärkeys piti kastepistettä, ilman lämpötilaa, korkeutta ja Ruka, Val di Fiemme ja Hochfilzen testipaikkoja. Permutaatiopohjaisen muuttujien tärkeys oli tosin muutama sadasosa parhaimman ja huonoimman muuttujan välillä. Se ei ole yllätys mallin huonon ennustuskyvyn takia.

Satunnaismetsän makrokeskiarvoistettu sisäinen tarkkuus on 11.8 %, makrokeskiarvoistettu herkkyys 12.3 % ja makrokeskiarvoistettu F1-arvo 11.1 %. Luokkakohtaiset sisäiset tarkkuudet kolmelle yleisimmälle voidemerkille testiaineistossa olivat 17 % REX-voidemerkille, 20 % STAR-voidemerkille ja noin 17 % TOKO-voidemerkille. Luokkakohtaiset herkkyydet noin 19 % REX-voidemerkille, 20 % STAR-voidemerkille ja 15 % TOKO-voidemerkille. Parhain luokkakohtainen F1-arvo 28 % tuli SWIX-voidemerkille, joita ennustettavana testiaineistossa on 14 näytettä. SWIX-voidemerkin sisäinen tarkkuus oli 26.7 % ja herkkyys 28.6 %.

5.6.2 Neuroverkko

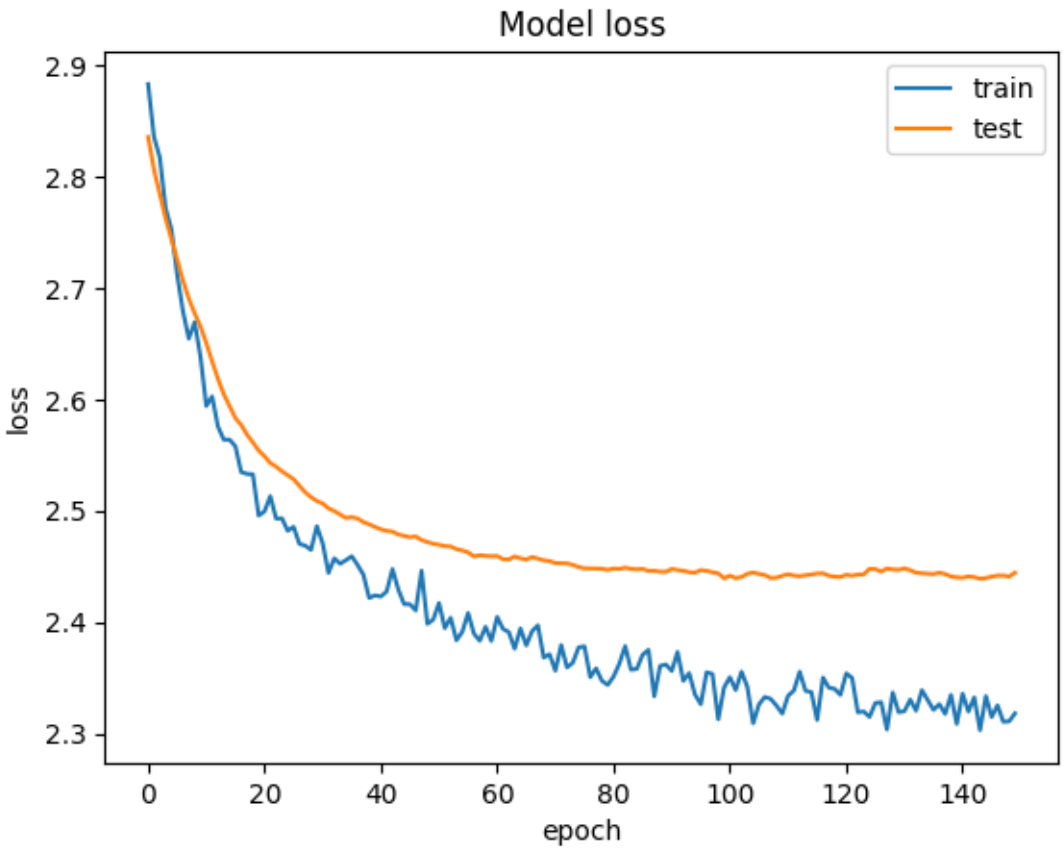
MLP-luokittelijamallin makrokeskiarvoiksi tuli sisäiselle tarkkuudelle 13.7 %, herkkyydeksi 13.7 % ja F1-arvoksi 12.4 %. Tarkkuudeksi malli sai noin 18.4 %. Kolmelle yleisimmälle voidemerkille luokkakohtaiset sisäiset tarkkuudet ovat 19 % REX-voidemerkille, 20 % STAR-voidemerkille ja 20.1 % TOKO-voidemerkille. Herkkyydet olivat 29.6 % REX-voidemerkille, 20 % STAR-voidemerkille ja 25 % TOKO-voidemerkille. Myös MLP-luokittelija sai parhaan F1-arvon SWIX-voidemerkille 34.8 %. jossa sisäinen tarkkuus oli 44.4 % ja herkkyys 28.6 %.

Kuvassa 9. on TensorFlow-kirjaston neuroverkkomallin tarkkuus opetuksen aikana opetus- ja testiaineistolle jokaiselta opetuskierrokselta.



Kuva 9. Neuroverkon tarkkuus opetuskierroksittain

Kuvassa 10. on TensorFlow-kirjaston neuroverkkomallin tappio opetus- ja testiaineistolle jokaiselta opetuskierrokselta.



Kuva 10. Neuroverkon häviö opetuskierroksittain

Kuvien 9. ja 10. perusteella voimme havaita, että malli on alkanut hieman ylisovittumaan opetusaineistoon. Malli on siis alkanut oppimaan opetusaineistosta kohinaa tai muuta satunnaista vaihtelua. Yksi syy mallin ylisovittumiseen voi olla liian vähäinen opetusaineiston määrä. [45].

TensorFlow-kirjaston neuroverkon sisäisten tarkkuuksien makrokeskiarvo on 11.7 %, herkkyyden makrokeskiarvo 12 % ja makrokeskiarvo F1-arvolle on 9.2 %. Luokkakohtaiset metriikat voidemerkeille on REX-voidemerkille 15.4 % sisäiseksi tarkkuudeksi ja 22.2 % herkkyydeksi; STAR-voidemerkille 20 % sisäiseksi tarkkuudeksi ja 36 % herkkyydeksi; TOKO-voidemerkille 21.7 % sisäiseksi tarkkuudeksi ja 25 % herkkyydeksi. Paras luokkakohtainen F1-arvo 33.3 % tuli SWIX-voidemerkille, missä sisäinen tarkkuus on 31.3 % ja herkkyys 35.7 %. Neuroverkon evaluate -menetelmän palauttama tarkkuus on noin 19.3 %.

5.6.3 Yhteenveto tuloksista

Taulukossa 3. on esitetty kehitettyjen koneoppimismallien tuloksia luvussa 5.2.1 valituille metriikoille.

Taulukko 3. Taulukko ennustusmallien metriikoiden tuloksista

Malli	**Tarkkuus**	**Sisäinen tarkkuus (makrokeskiarvo)**	**Herkkyys (makrokeskiarvo)**	**F1-arvo (makrokeskiarvo)**
Dummy CLF	13 %	0.7 %	5.5 %	1.3 %
Satunnaismetsä	14.5 %	11.8 %	12.3 %	11.1 %
MLP	18.4 %	**13.7 %**	**13.7 %**	**12.4 %**
TF: Neuroverkko	**19.3 %**	11.7 %	12 %	9.2 %

Verratuista malleista parhaan tarkkuuden sai TensorFlow-neuroverkko-luokittelija, parhaan makrokeskiarvoistetun sisäisen tarkkuuden sai MLP-luokittelija, parhaan makrokeskiarvoistetun herkkyyden sai MLP-luokittelija ja parhaan makrokeskiarvoistetun F1-arvon sai MLP-luokittelija.

Parhaan tarkkuuden verratuista malleista on saanut TensorFlow-kirjastolla kehitetty luokittelija. Mallin tarkkuus on 19.3 % tähän on vaikuttanut etenkin se, että kahdeksalle suosituimmalle voidemerkille on saatu keskimäärin noin 26 % tarkkuus. Joista parhaan noin 67 % tarkkuuden on saanut MAPLUS-voidemerkki. Muista verratuista malleista MLP-luokittelijan 18.4 % tarkkuuteen on vaikuttanut etenkin se, että kolmelle suosituimmalle voidemerkille on ennustettu noin 20 % oikein.

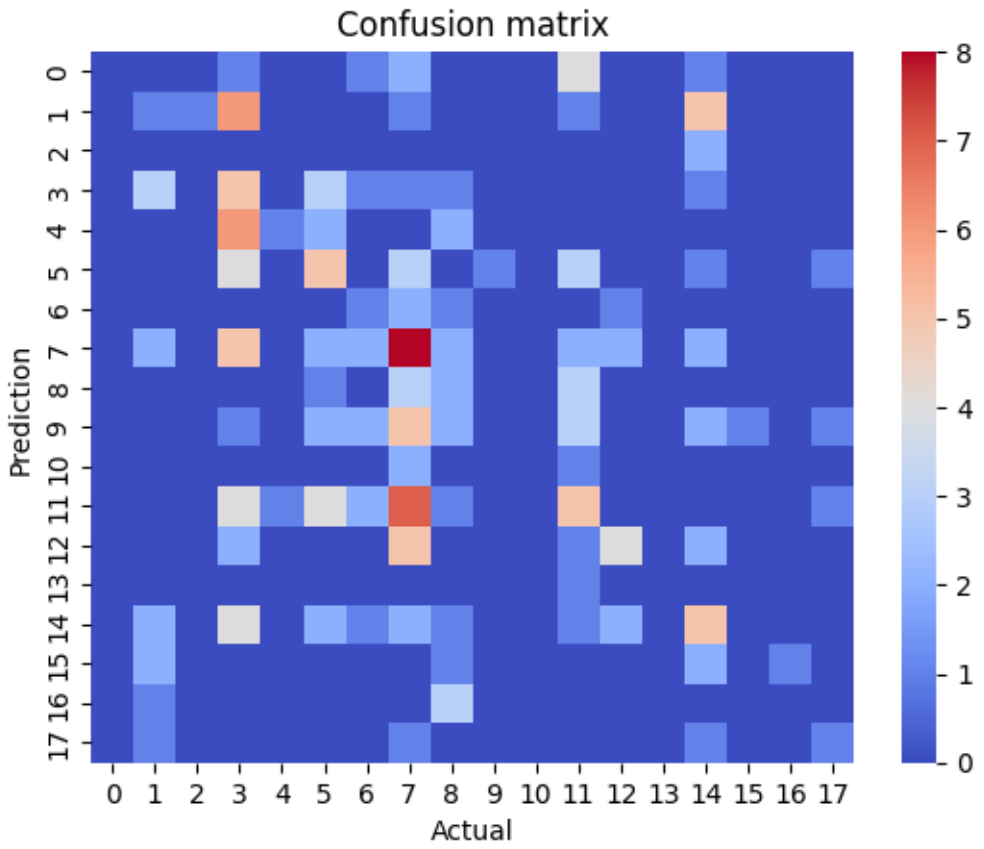
Parhaan makrokeskiarvoistetun sisäisen tarkkuuden sai MLP-luokittelija. MLP-luokittelijassa makrokeskiarvoitettua sisäistä tarkkuutta nosti etenkin HWK-voidemerkin 50 % sisäinen tarkkuus, sekä SWIX-voidemerkin 44 % sisäinen tarkkuus ja ZIPP-voidemerkin 25 % sisäinen tarkkuus. Vaikka Tensorflow-kirjaston neuroverkon tarkkuus on ollut verratuista malleista paras, on se saanut makrokeskiarvoitetuksi sisäiseksi tarkkuudeksi vain 11.7 %. Tähän on etenkin vaikuttanut se, että vain kolmelle voiteelle on saatu muita selkeästi korkeampi sisäinen tarkkuus, jotka ovat noin 67 % MAPLUS-voidemerkille, noin 40 % GALLIUM-voidemerkille ja noin 31 % SWIX-voidemerkille.

Muille 4 ennustetulle voidemerkille sisäinen tarkkuus on noin 15 % ja 20 % välillä. Lopuille 11 voidemerkille 0 %.

Parhaan makrokeskiarvoistetun herkkyyden sai MLP-luokittelija. MLP-luokittelijassa suureen herkkyyteen vaikutti etenkin HOLMENKOL-voidemerkin 33 % herkkyys, REX-voidemerkin noin 30 % herkkyys ja SWIX-voidemerkin noin 29 % herkkyys. Muille ennustetuille voidemerkeille on myös tullut yli 20 % herkkyys, paitsi HWK- ja GALLIUM-voidemerkeille, joiden herkkyys on alle 10 %. Toiseksi parhaan makrokeskiarvoistetun herkkyyden sai satunnaismetsä. Satunnaismetsä on pyrkinyt ennustamaan yhtä montaa voidemerkkiä kuin MLP-luokittelija, mutta ei ole saanut niin korkeita herkkyyksiä voidemerkeille.

Mikään verratuista malleista ei osannut ennustaa SKIGO-voidemerkkiä. Vaikka se on ollut kuudenneksi yleisin voidemerkki 18 ennustettavasta voidemerkistä. Tähän saattaa vaikuttaa osittain se, että SKIGO-voidemerkkiin liittyvien näytteiden muuttujien arvot eivät ole olleet selkeästi korkeita tai matalia. Eikä SKIGO-voidemerkkiin liittyvät näytteet ole olleet selkeästi mihinkään tiettyyn testipaikkaan keskittyneitä. SWIX-voidemerkki on ollut yksi parhaiten ennustettavissa oleva voidemerkki kakkien verrattavien mallien keskuudessa. SWIX-voidemerkin ennustuskykyyn on saattanut vaikuttaa se, että se on toiminut muita voidemerkkejä paremmin etenkin yhdessä suositussa testipaikassa.

Kuvassa 11. on taulukko MLP-mallin sekaannusmatriisista.



Kuva 11. MLP-mallin ennustuksen sekaannusmatriisi

Kuvasta 11. voimme havaita, että ennustettavalle luokalle 7 on tullut määrällisesti eniten oikein ennustettuja näytteitä. Luokalle 7 on myös testiaineistossa eniten ennustettavia näytteitä. Luokka 7 on REX-voidemerkki. REX-voidemerkin väärin ennustetuista näytteistä suurin osa on mennyt luokalle 11. Luokka 11 on STAR-voidemerkki, mikä on testiaineistossa toiseksi eniten esiintyvä voidemerkki. STAR-voidemerkkiä on ennustettu jopa enemmän REX-voidemerkille, mitä STAR-voidemerkille. Luokkia 0, 10 ja 13 ei ole ennustettu kertaakaan. Voidemerkkeinä nämä ovat BRIKO, jonka näytteitä kokoaineistosta löytyy 30, SR20, jonka näytteitä kokoaineistosta löytyy 8 ja TESTLGRS, jonka näytteitä kokoaineistosta löytyy 7. Muita voidemerkkejä, joita ei ole testiaineistosta kertaakaan ennustettu oikein on VOR, VAUHTI, HELXB ja SKIGO. Näissä kaikissa paitsi SKIGO-voidemerkissä yhtenäisenä tekijänä on vähäinen näytteiden määrä aineistossa. Parhaat ennustus tulokset vähäisen näytemäärän voidemerkistä on saanut ZIPP-voidemerkki. Kokoaineistossa voidemerkkiä on 5. vähiten, mutta testiaineiston 4 näytteestä 1 on onnistuttu ennustamaan oikein. Sekaannus matriisissa ZIPP-voidemerkki on luokka 17.

6 Pohdinta

Opinnäytetyössä havaittiin siis, että hiihdon luisto-ominaisuuksiin vaikuttaa moni asia. Kuten painovoima, kitka, lumi, ilmanvastus, voitelu ja kuviointi sekä konekuviointi että käsikuviointi. Yhteenveto kehitettyjen koneoppimismallien tuloksista on se, että mikään kehitetty koneoppimismalli ei osannut suoraan ennustaa hyväksi merkattua voidetta niin tarkasti, että niitä pystyttäisiin suoraan hyödyntämään. Osan voidemerkkien ennustuskyky oli parempi, kun taas osaa voidemerkeistä ei osattu ennustaa ollenkaan.

Kuten luvussa 5.1.1 aineiston esittely on esitelty, että alkuperäisen aineiston näytteiden määrä on noin 9000 riviä, mistä voidetestejä oli noin 3000 näyteriviä. Näistä 3000 rivistä luvun 5.1.2 aineiston esikäsittely -toimintojen jälkeen algoritmille hyödyllisiä näyterivejä jäi vain noin 1000 näytettä. Yksi syy kehitettyjen mallien huonoihin ennustettavuuskykyihin on liian vähäisen opetusaineiston määrä.

Toinen voiteiden huonoon ennustettavuuteen vaikuttava tekijä voi olla se, että voiteet ovat liian samanlaisia. Vaikka malli ei ennustuksen mukaan suositellut testiaineistossa oikeaksi merkattua voidetta, ei se silti tarkoita, etteikö voide olisi hyvä. Sillä useampi voide voi olla hyvä samanlaiselle keliolosuhteelle. Jos luokittelumallien tarkkuuden sijasta käyttäisimme mittaria, joka kertoisi suositellun voiteen paremmuutta oikeaksi merkatusta voiteesta. Voisimme mahdollisesti saada paremmin tietoa siitä, onko jokin malli toiminut hyvin tai ei. Tällaiseen mallin kouluttamiseen on olemassa cost-sensitive learning -menetelmiä, mutta tällaisen menetelmän hyödyntämiseen tarvittaisiin toki ”hinta” väärin ennustetulle voiteelle eli tieto siitä, kuinka lähellä ennustettu voide on testiin merkatusta voiteesta. TensorFlow-neuroverkkomallin ulostulokerroksen aktivointifunktiona on käytetty softmaxia. Softmax-aktivointifunktiosta on mahdollista nähdä, mitä neuroverkko on ennustanut toiseksi parhaaksi tai kolmanneksi parhaaksi voiteeksi, vaikka tätä tietoa ei ole hyödynnetty opinnäytetyössä. Voisi näitä voiteita myös verrata siihen, ovatko ne testissä hyväksi merkattu voide.

Suosittelualgoritmin toimivuudesta voimme esimerkkinä testata MLP-mallia opinnäytetyön kirjoitushetkellä oleviin keliolosuhteisiin Vuokatissa, jotka osan saamme Ilmatieteen laitoksen sivuilta ja verrata näitä voidetietokannassa oleviin testeihin. Ilmatieteen laitoksen verkkosivut kertovat keliolosuhteiden Sotkamossa olevan -15 celsiusasteen ilmanlämpötila, suhteellisen kosteuden olevan 88 % ja kastepisteen olevan -16.6 celsiusastetta. Lumen lämpötila 1 cm:n syvyydessä-

ja uran kosteus -muuttujien arvot syötetään oman arvion mukaan ja suksen tunnisteeksi syötetään maastohiihto. Näillä muuttujien arvoilla malli ennustaa voidemerkiksi REX-voidemerkin, mikä on myös ollut vastaavilla keliolosuhteilla Falunissa käytetty voiteen voidemerkki. Kyseisessä olosuhteessa testattu REX-voidemerkin voide oli TK83-voidetta, joka omassa testipakassa oli sijalla 5.

Luvussa 3.5 Voitelu on mainittu hiihdossa kansallisella ja kansainvälisellä tasolla tulevasta fluorikiellosta. Fluorin käyttäminen voiteissa parantaa merkittävästi suksien luisto-ominaisuuksia. Vaikka fluorikielto on tulossa voimaan, se ei vaikuta tässä opinnäytetyössä kehitetyn mallin ennustuskykyyn tässä opinnäytetyössä käytetylle aineistolle.

Kuten luvussa 3.3 on todettu, että sana lumi kattaa useamman tyyppisiä lumenrakenteita. Tämän takia uuteen voidetestitietokannasta ladattavaan raporttiin olisi hyvä lisätä muuttuja lumenrakenteen tyypistä. Lisäksi, koska voiteita oli vertailtu pienempien suksitestipakkojen sisällä keskenään, pitäisi raporttiin myös lisätä testipakkojen ID. Jonka avulla testipakat olisi helppo tunnistaa toisistaan.

Ihannetilanteessa opetusaineisto olisi ollut sellainen, missä kaikkia voiteita on testattu, kaikissa testipaikoissa, kaikilla suksilla ja samanlaisilla keliolosuhteilla, minkä myötä algoritmi olisi opetettu. Ihannetilanteessa, jos joku voiteiden asiantuntija olisi antanut myös omat suosittelunsa kaikille keliolosuhteille tällaisen täydellisen aineiston pohjalta, olisi sitä sitten voitu verrata kehitettyyn algoritmiin.

Yksi merkittävä lisä tulevaan voidetestitietokannan käyttöliittymään testien lisäämiseen voisi olla myös valmiiksi kirjoitetut voiteiden nimet. Sillä nykyisessä aineistossa pelkästään sama voidemerkki on voitu kirjoittaa viidellä eri tavalla. Uusien voiteiden lisääminen voidetietokantaan voisi tapahtuma pääkäyttäjän tai vastaavien muokkausominaisuuksien omaavan toimesta. Voiteet voi sitten testien tekijät valita testejä syöttäessä valmiista valikosta. Tämän avulla vältetään paremmin virheellisesti nimettyjen voiteiden määrää. Niin kuin voiteiden kirjoittamiselle pitää kehittää selkeämpi merkkaustapa testeihin, niin myös voideseoksille, jotka on tehty useammasta voiteesta jollain tietyllä suhteella. Sillä tällä hetkellä nykyisessä aineistossa tällaisten voideseosten merkintätapa vaihtelee testien välillä.

Tässä opinnäytetyössä tehtyä työtä jatketaan ja hyödynnetään suksihuollonkehityshankkeessa. Esimerkiksi uusien testien, suksien, voiteiden, käsi- ja konekuvioinnin lisäys sovellusalustan kehi-

tyksessä. Sovelluksen kehityksessä pyritään huomioimaan tämän opinnäytetyön aikana huomioidut sinne mahdollisesti lisättävät ominaisuudet. Sekä sovelluksen kanssa toimivassa tietokantarakenteessa.

Lähteet

1 Suksihuollon kehittämishanke hankekuvaus, Jyväskylän yliopiston Vuokatin liikuntateknologian yksikkö, [Viitattu 12.1.2024], Saatavilla: https://www.jyu.fi/fi/hankkeet/suksihuollon-kehittamishanke-olympiavalmennuskeskus-vuokatti-rukan-lajeissa

2 Perinteisen hiihtotavan simulointi suksen liikutuslaitteessa, sekä luistoon ja pito-ominaisuuksiin vaikuttavat tekijät, Veli Kolehmainen, 2006, Saatavilla: http://urn.fi/URN:NBN:fi:jyu-20094151454

3 Koneoppiminen, Jeremias Penttilä, 2015, Saatavilla: http://www.urn.fi/URN:NBN:fi:jyu-201603211906

4 Kirvesniemi H. Hyvä hiihtokoulu, Teos, 2006

5 Kisatason luistelusukset testissä, Anttila S. Pentsinen A., 2021, [Verkkoartikkeli], [Viitattu 12.9.2023] Saatavilla: https://juoksija.fi/hiihto/hiihtovarusteet/kisatason-luistelusukset-testissa-ohjeet-suksen-valintaan/

6 Fysiikan oppikirja/Liike ja voima, Wikikirjasto, 2018, [Internet], [Viitattu 26.10.2023] Saatavilla: https://fi.wikibooks.org/wiki/Fysiikan_oppikirja/Liike_ja_voima

7 Kitka, Wikipedia, [Internet], [Viitattu 26.10.2023] Saatavilla https://fi.wikipedia.org/wiki/Kitka

8 Suksen pito- ja luisto-ominaisuuksien muutoksen vaikutus voimantuottoon ja lihasaktiivisuuteen maksimaalisessa pitkäkestoisessa hiihtosuorituksessa perinteisellä hiihtotavalla, Veli-Matti Nieminen, 2013, Saatavilla: http://www.urn.fi/URN:NBN:fi:jyu-201302201250

9 Kitka, Wikipedia, [Valokuva]. [Viitattu 4.3.2023]. Saatavilla: https://commons.wikimedia.org/wiki/File:Friction_alt.svg

10 Ilmanvastus, Wikipedia, [Internet], [Viitattu 26.10.2023] Saatavilla: https://fi.wikipedia.org/wiki/Ilmanvastus

11 Luistovoiteluohjeet, Rex, 2022, [Verkkoartikkeli], [Viitattu 21.9.2023] Saatavilla: https://rex.fi/tietopankki/luistovoiteluohjeet

12 Suksien voitelu – helpot ohjeet aloittelijoille, [Verkkoartikkeli], [Viitattu 21.9.2023] Saatavilla: https://retkivinkit.fi/suksien-voitelu-helpot-ohjeet-aloittelijalle/

13 Kolmen vuoden vatuloinnista tuli miljoonien lasku – hiihdon fluorikielto ei ole edes voimassa, mutta se on maksanut Suomellekin jo satojatuhansia, Husu A., 2022, [Verkkouutinen], [Viitattu 21.9.2023] Saatavilla: https://yle.fi/a/74-20004308

14 FIS to fully implement fluor wax ban at start of 2023-24 season, FIS, 2023, [Verkkoartikkeli], [Viitattu 21.9.2023] Saatavilla: https://www.fis-

ski.com/en/international-ski-federation/news-multimedia/news-2022/fis-to-fully-implement-fluor-wax-ban-at-start-of-2023-24-season

15 Kansallista fluorikieltoa valmistellaan kaudelle 2024–2025, Hiihtoliitto, [Verkkoartikkeli], [Viitattu 21.9.2023] Saatavilla: https://hiihtoliitto.fi/kansallista-fluorikieltoa-valmistellaan-kaudelle-2024-2025/

16 Muovi-Set Finland Oy, Speedy ski roller ja SSR toimintaperiaate, [internet], [viitattu 20.9.2023], Saatavilla: http://www.skiroller.fi/index.php?id=11

17 Maastohiihtosuksen pohjakuvioinnin mittauslaitteen suunnittelu, Juho Salmela, 2022, Saatavilla: http://www.urn.fi/URN:NBN:fi:amk-2022060917105

18 Some aspects of ski base sliding friction and ski base structure, Dag Anders Moldestad, 1999, Saatavilla: http://hdl.handle.net/11250/236372

19 Koneoppiminen : ohjattu oppiminen taloustieteellisessä tutkimuksessa, Jukka-Pekka Jauhiainen, 2019, Saatavilla: http://urn.fi/URN:NBN:fi:oulu-201906052387

20 Artificial Neural Networks and its Applications, GeeksforGeeks, [Valokuva], [Viitattu 4.10.2023], Saatavilla: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

21 Koneoppiminen vedonlyönnissä, Oliver Saarela, 2022, Saatavilla: http://www.urn.fi/URN:NBN:fi:amk-2022060114087

22 Random Forests, Medium, Yehoshua R., 2023, [Valokuva], [Viitattu 4.10.2023] Saatavilla: https://medium.com/@roiyeho/random-forests-98892261dc49

23 Cambridge University Press, Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz Shai Ben-David, 2014

24 Beamforming analysis using Random Forest classifier, Riikka Valkama, 2021, Saatavilla: http://urn.fi/URN:NBN:fi:oulu-202105208048

25 Accuracy, precision, and recall in multi-class classification, Evidently AI Team, [Internet], [Viitattu 16.11.2023], Saatavilla: https://www.evidentlyai.com/classification-metrics/multi-class-metrics

26 Koneoppimismallien käyttö luokittelutehtävässä, Harri Huikuri, 2022, Saatavilla: https://urn.fi/URN:NBN:fi:amk-2022101321166

27 Macro F1 and Macro F1, Juri Opitz, Sebastian Burst, 2021, Saatavilla: https://doi.org/10.48550/arXiv.1911.03347

28 About Us, Jupyter, [Internet], [Viitattu 9.11.2023], Saatavilla: https://jupyter.org/about

29 Vesitutkimustulosten tulkinta, WatMan, 2010, [Verkkoartikkeli], [Viitattu 26.9.2023], Saatavilla: https://www.watman.fi/pdf/vedenlaatu.pdf

30 Principal component analysis (PCA), Scikit-Learn, [Internet], [Viitattu 23.11.2023], Saatavilla: https://scikit-learn.org/stable/modules/decomposition.html#pca

31 Ways to Detect and Remove the Outliers, Natasha Sharma, 2018, [Verkkoartikkeli], [Viitattu 9.10.2023], Saatavilla: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

32 Why “1.5” in IQR Method of Outlier Detection?, Shivam Chaudhary, 2019, [Verkkoartikkeli], [Viitattu 10.10.2023], Saatavilla: https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

33 Feature Engineering: Scaling, Normalization, and Standardization, GeeksforGeeks, [Internet], [Viitattu 10.10.2023], Saatavilla: https://www.geeksforgeeks.org/ml-feature-scaling-part-2/

34 Using Correlation Coefficients in Research Papers, Editor World, [Internet], [Viitattu 19.10.2023], Saatavilla: https://www.editorworld.com/article/using-correlation-coefficients-in-research-papers

35 Top Techniques to Handle Missing Values Every Data Scientist Should Know, Zoumana Kelta, 2023, [Internet], [Viitattu 24.10.2023], Saatavilla: https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values

36 What Is Data Preparation in a Machine Learning Project, Jason Brownlee, 2020, [Internet], [Viitattu 10.10.2023], Saatavilla: https://machinelearningmastery.com/what-is-data-preparation-in-machine-learning/

37 Categorical Data Encoding Techniques, Krishnakanth Naik Jarapala, 2023, [Internet], [Viitattu 16.10.2023], Saatavilla: https://medium.com/aiskunks/categorical-data-encoding-techniques-d6296697a40f

38 Sklearn Neural Network MLPClassifier, Scikit-learn, [Internet], [Viitattu 22.11.2023], Saatavilla: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

39 A complete Understanding of Dense Layers in Neural Network, Yugesh Verma, 2021, [Verkkoartikkeli], [Viitattu 23.11.2023] Saatavilla: https://analyticsindiamag.com/a-complete-understanding-of-dense-layers-in-neural-networks/

40 A Gentle Introduction to the Rectified Linear Unit (ReLU), Jason Bronwlee, 2020, [Internet], [Viitattu 23.11.2023] Saatavilla: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

41 A Gentle Introduction to Dropout for Regularizing Deep Neural Networks, Jason Brownlee, 2019, [Internet], [Viitattu 23.11.2023], Saatavilla: https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/

42 How to Choose an Activation Function for Deep Learning, Jason Brownlee, 2021, [Internet], [Viitattu 18.1.2024], Saatavilla: https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/

43 Feature importances with a forest of trees, Scikit-learn, [Internet], [Viitattu 9.11.2023], Saatavilla: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

44 Permutation Importance vs Random Forest Feature Importance (MDI), Scikit-learn, [Internet], [Viitattu 9.11.2023], Saatavilla: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html

45 How to use Learning Curves to Diagnose Machine Learning Model Performance, Jason Brownlee, 2019, [Internet], [Viitattu 28.11.2023], Saatavilla: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Liitteet

Liite 1.

Lähdekoodi Ura-muuttujan imputointiin

```
# replacing missing values in ura

mice_kernel = ImputationKernel(

   data = dataset[['Ura', 'LI', 'L1', 'KP']],

   random_state=42,

   imputation_order=['Ura']

)

mice_kernel.mice(2)

mice_imputation = mice_kernel.complete_data()

dataset.loc[:, 'Ura'] = mice_imputation['Ura']
```

Liite 2.

Lähdekoodit opinnäytetyössä käytettyjen mallien tekemiseen

```
# Dummy classifier

from sklearn.dummy import DummyClassifier


dummy_clf = DummyClassifier(strategy='most_frequent')

dummy_clf.fit(X_train, y_train)


# Random forest classifier

from sklearn.ensemble import RandomForestClassifier


clf = RandomForestClassifier(random_state=42)

clf.fit(X_train, y_train)


# MLP classifier

from sklearn.neural_network import MLPClassifier


nn_clf = MLPClassifier(random_state=42, max_iter=10_000, hidden_layer_sizes=(100,))

nn_clf.fit(X_train, y_train)


# Tensorflow neural network
```

```
from tensorflow import keras


model = keras.Sequential([
    keras.layers.Input(shape=(X_test.shape[1])),
    keras.layers.Dense(16, activation='relu'),
    keras.layers.Dropout(0.5),
    keras.layers.Dense(dataset['Voide 1'].nunique(), activation='sigmoid')
])


model.compile(optimizer='adam',
              loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
             metrics=[keras.metrics.SparseCategoricalAccuracy()])


history = model.fit(
    X_train.to_numpy().astype('float32'),
    y_train, epochs=150,
    validation_data=(X_test.to_numpy().astype('float32'), y_test),
    batch_size=32,
    verbose=0
)
```