Tampereen yliopisto

Tuomas Nuutila

SELITETTÄVÄ TEKOÄLY TALVIMEREN-KULUSSA

Diplomityö
Johtamisen ja talouden tiedekunta
Tarkastajat: Heikki Liimatainen
Jukka Huhtamäki
05/2022

TIIVISTELMÄ

Tuomas Nuutila: Selitettävä tekoäly talvimerenkulussa
Diplomityö
Tampereen yliopisto
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Toukokuu 2022

Tekoälyn hyödyntäminen ja erilaisten tekoälyjärjestelmien käyttöönotto voimistuu vuosi vuodelta yhä enemmän ja siten on vaarana niin sanottujen läpinäkymättömien ”mustan laatikon” tekoälymallien lisääntyminen. Tällaisten mallien kohdalla ei pystytä suoraan selittämään miksi ja mihin malli perustaa toimintansa. Tämän vuoksi selitettävän tekoälyn (XAI) tärkeys on ymmärretty enenevissä määrin, jotta läpinäkymättömyyden ongelmaan pystyttäisiin vastaamaan. Itämeri on yksi harvoista maailman meristä, jossa tapahtuu jäänmurtoa. Talvimerenkulun merkitys Suomen kaupalle on merkittävä ja viivästykset voivatkin aiheuttaa merkittäviä tappioita. Jää luo suuren haasteen niin kulkemiselle, kuin myös tekoälyratkaisujen kehittämiselle ja niiden käyttämiselle, sillä jäätilanteet voivat muuttua hyvinkin nopeasti, eikä sen tarkalle ja riittävän nopealle mallintamiselle ole vielä kehitetty keinoja.

Tämän diplomityön tarkoituksena on tarkastella selitettävää tekoälyä talvimerenkulussa. Työssä käydään läpi teoriaa tekoälystä ja laivaliikenteestä sekä selitettävästä tekoälystä, jonka jälkeen analysoidaan työn prosessia. Selitettävyydellä on selkeitä hyötyjä, kuten luottamuksen, läpinäkyvyyden sekä ymmärryksen luonti unohtamatta regulaatioiden ja lakien noudattamista, joita muun muassa EU:n tasolla on tulevaisuudessa mahdollisesti tulossa. Selitettävyys tulisi myös ymmärtää osana koko prosessia, eikä vain mallin selitteenä loppukäyttäjälle. Mallin kehitysprosessissa esimerkiksi erilaiset SHAP- ja LIME-työkalut ovat hyödyllisiä selittäjiä kehittäjille, jotka tarvitsevat syvempää tietoa muuttujista ja niiden vaikutuksesta malliin. Tällaiset ovat kuitenkin usein turhan epäselkeitä ja vaikeasti tulkittavia loppukäyttäjille, jolloin heille tulisi tarjota selite, joka esimerkiksi selkeästi tekstiformaatissa kertoo mallin tekemästä päätöksestä tai suosituksesta.

Työn tuloksena syntyi hahmotelma tulevaisuuden XAI-ratkaisusta talvimerenkulussa. Selitteiden pohjana käytettiin tämän työn ulkopuolella kehitettyä Random forest -algoritmiin perustuvaa koneoppimismallia, jota selitettiin SHAP-työkalulla. Lisäksi Tableau-visualisointien avulla tutkittiin dataa.

Avainsanat: tekoäly, selitettävä tekoäly, talvimerenkulku

ABSTRACT

Tuomas Nuutila: Explainable AI in Winter navigation
Master of Science Thesis
Tampere University
Master's Degree Programme in Information and Knowledge Management
May 2022

The use of Artificial Intelligence systems is growing year by year, and thus the growth of so-called opaque "black box" models may cause increased risk. In the case of such models, it is not possible to directly explain why and where the model is based. Therefore, the importance of Explainable Artificial Intelligence (XAI) has been increasingly understood to address the problem of opacity. The Baltic Sea is one of the few seas in the world where icebreaking takes place. The importance of Winter Navigation to Finnish trade is significant and delays can cause significant losses. Ice poses a major challenge for vessel navigation and also to the development and use of Artificial Intelligence solutions as the ice situations can change very quickly and there has not yet been found solution to model the ice accurately and quickly enough.

The purpose of this thesis is to examine Explainable Artificial Intelligence in Winter Navigation. The thesis goes through the theory of Artificial Intelligence, shipping, and Explainable Artificial Intelligence, followed by an analysis. Explanability has clear benefits, such as building trust, transparency and understanding. Also, following the regulations and laws that may come into play in the future, for instance at EU level. Explanability should also be understood as part of the whole process, and not just as an explanation of the model to the end user. In the model development process, for example, various SHAP and LIME tools are useful explanators for developers who need more in-depth information about variables and their impact on the model. However, these kinds of explanations are often too difficult to interpret for end-users, in which case they should be provided with an explanation that, for example, clearly indicates in a text format the decision or recommendation made by the model.

The work resulted in a draft of the future XAI solution in the Winter Navigation. The explanations were based on a machine learning model based on the Random Forest algorithm developed outside of this work, which was explained with the SHAP tool. In addition, data were examined using Tableau visualizations.

Keywords: Artificial Intelligence, Explainable AI, Winter navigation

ALKUSANAT

Keväällä 2021 sain mahdollisuuden päästä kirjoittamaan diplomityötä aiheesta, joka oli itselleni vieras, mutta samalla hyvin kiehtova. Nyt vuotta myöhemmin voi todeta, että prosessi on ollut pitkä ja opettavainen sekä aiheena edelleen kiehtova. Näitä sanoja kirjoittaessa alkaa vihdoin ymmärtämään, että työ on valmis ja samalla myös yksi merkittävä vaihe elämästä lähestyy loppuaan valmistumisen myötä.

Haluan kiittää tasapuolisesti kaikkia, jotka ovat olleet mukana auttamassa matkan varrella. Erityisesti haluan kiittää Timo Lehtosta, jonka tuki on ollut arvokasta heti työn alkumetreiltä lähtien. Lisäksi haluan kiittää Solitaa, Väylävirastoa sekä AIGA-tutkimushanketta työn mahdollistajana. Suuri kiitos myös Tampereen yliopiston Heikki Liimataiselle sekä Jukka Huhtamäelle ohjauksesta ja avusta läpi prosessin. Viimeisimpänä, muttei vähäisimpänä haluan vielä kiittää ystäviäni sekä perhettäni ja muita läheisiä, joiden tuesta olen saanut energiaa kirjoittaa työtä.

Tampereella, 18.5.2022

Tuomas Nuutila

SISÄLLYSLUETTELO

LYHENTEET JA MERKINNÄT

XAI	Explainable Artificial Intelligence (Selitettävä tekoäly)

1. JOHDANTO

Tässä luvussa käydään läpi yleisesti työn taustaa ja tavoitteita sekä sen rajausta, tutkimuskysymyksiä ja rakennetta. Lisäksi perehdytään työssä käytettyyn tutkimuksen viitekehykseen hyödyntäen Saunders et al. (2009) kehittämää sipulimallia.

1.1 Talvimerenkulku ja selitettävä tekoäly

Laivojen kulku- ja sijaintitiedot jääolosuhteissa ovat tärkeässä osassa toimitusketjujen toiminnan kannalta. Muun muassa Suomenlahdella ja Pohjanlahdella jäät aiheuttavat laivojen juuttumista ja odottelua, mikä voi viivästyttää merkittävästi laivojen satamaan saapumista. Suomessa noin 90 % viennistä ja 80 % tuonnista kulkee meriteitse, joten viivästymiset voivat aiheuttaa merkittäviä lisäkustannuksia heikentäen ulkomaankaupan hintakilpailukykyään. Luotettavat, turvalliset sekä sujuvat kulkuyhteydet ovat Suomen elinkeinoelämän ja siten koko yhteiskunnan kilpailukyvyn elinehto. Suomen ilmaston ja maantieteellisen sijainnin vuoksi kaikki rannikon satamat jäätyvät talvisin, joten jäänmurto on välttämätöntä. (Ojala et al., 2020; Liikenne- ja viestintäministeriö, 2014; Toivola, 2016)

Itämeri on yksi harvoista maailman meristä, jossa joudutaan käyttämään jäänmurtajia merenkulun ylläpitämiseksi. Talvimerenkulun ylläpitäminen on hyvin tärkeää, sillä yli 85 % Suomen bruttokansantuotteen muodostavasta ulkomaankaupasta tulee merikuljetuksista ja ilman kunnollista jäänmurtoa kaikki Suomen satamat jäätyvät normaalitalvena, jolloin rahdin toimitus olisi talvella mahdotonta meriteitse (Toivola, 2016). Ilmatieteenlaitoksen (n.d.) mukaan Itämeren jäätyminen alkaa Perämeren pohjoisosista ja Suomenlahden pohjalta tyypillisesti marraskuun aikana ja on laajimmillaan tammikuun ja maaliskuun aikana. Tällöin keskimäärin noin kaksi viidesosaa Itämerestä eli noin 170 000 km^2 on jääpeitteen alla, mutta esimerkiksi vuonna 2020 jää peitti vain noin 37 000 km^2 alueen.

Talvet voidaan luokitella neljään eri ankaruusluokkaan: leuto, keskimääräinen, ankara sekä erittäin ankara. Talvet, jolloin jääpeite on alle 115 000 km^2 ovat ankaruusluokaltaan leutoja, yli 230 000 km^2 ankaria ja erittäin ankaria, jos jääpeite on yli 345 000 km^2. Ankaruusluokat eivät kuitenkaan ota huomioon jäänpeitteen paksuutta, tiiviyttä tai ahtautumisastetta. Tärkeää on siten huomioida, että talvimerenkulun näkökulmasta leuto

talvi ei tarkoita välttämättä helppoa kulkua, eikä ankara vaikeaa kulkua, joskin avustusmatkat ovat pitkiä. Esimerkiksi leutoina talvina voi pakkaskelien välissä olevat lämpimät ja tuuliset kelit aiheuttaa jäiden liikkumista ja siten ahtautumista ja puristusta. Myös pakkasjaksot, jolloin tuuli on heikkoa, lisäävät jäiden määrää, jotka tuulen yltyessä ajautuvat kiinteän ajojääkentän reunaan muodostaen niin sanottuja sohjovöitä, jotka ovat talvimerenkulussa haastavia kulkea. (Ilmatieteenlaitos, n.d.)

Tekoälystä puhuttaessa selitettävyyden merkitys on tullut yhä tärkeämmäksi vuosi vuodelta, sillä tekoälymalleja ja niiden päätöksiä tulisi pystyä selittämään ja siten antamaan läpinäkyvyyttä mallille. Selitettävyys voi kuitenkin tarkoittaa eri asioita eri sidosryhmille, jolloin tulee miettiä, millaisia selitteitä sidosryhmille annetaan. (Preece et al. 2018) Esimerkiksi siinä missä tekoälymallien kehittäjille SHAP-selitteet ovat hyödyllisiä, ovat ne loppukäyttäjälle mahdollisesti liian vaikeasti tulkittavia. Tämän vuoksi loppukäyttäjälle, kuten laivan ohjaajalle, selitteet tulisi olla helposti ja nopeasti tulkittavissa. Tässä työssä käydään läpi selitettävyyden eri sidosryhmiä, erilaisia tekoälyselitteitä sekä teknologioita, joilla selitteitä voidaan luoda. Talvimerenkulku on jääolosuhteiden vuoksi haastavaa niin kulkemisen kuin tekoälymallien luomisen kannalta. Talvimerenkulun viranomaisen mukaan tarkkaa jäädataa ei ole saatavilla muun muassa jään nopean liikkumisen vuoksi, jolloin jäädataa käyttävien mallien selitettävyys korostuu, jotta voidaan paremmin ymmärtää jään vaikutusta mallin toimintaan.

1.2 Tutkimuksen tausta ja tavoitteet

Tutkimuksen tavoitteena on tarkastella tekoälyn selitettävyyttä talvimerenkulun näkökulmasta sekä selvittää alusten kulkua jääolosuhteissa kuvailevaa analytiikkaa hyödyntäen. Tutkimus alkoi tutustumalla saatavilla olevaan avoimeen dataan. Diplomityön kirjoittajalle talvimerenkulku ja laivaliikenteestä kerättävä AIS-data sekä selitettävä tekoäly eivät olleet entuudestaan tuttuja, joten tutkimuksessa oli lähdettävä liikkeelle talvimerenkulun perusasioista. Tämä diplomityö on osa kokonaisuutta, jossa tutkitaan tekoälyä talvimerenkulussa. Työssä tehtyjen selitteiden pohjana toimiva koneoppimismalli on kehitetty toisen diplomityön yhteydessä.

Tähän mennessä laivojen jääolosuhteissa kulkemisen ennustaminen on ollut hyvin vähäistä, eikä tekoälyn selitettävyyttä ole hyödynnetty ennustemallien kuvailemiseen. Talvimerenkulussa on suurin haaste tekoälymallien ja niiden selittämisen kannalta on jäätilanteen huomioiminen, sillä tänä päivänä ei ole vielä keinoa, jolla pystyttäisiin saamaan riittävän tarkkaa jääinformaatiota riittävän nopealla aikasyklillä. Esimerkiksi satelliittikuvista saatavat jäädatat eivät ainakaan nykyisellään anna riittävän tarkkaa ja no-

peaa tietoa vallitsevasta jäätilanteesta tarkan ennustemallin luomiseksi. Selitettävän tekoälyn sekä kuvailevan analytiikan avulla tutkimuksessa pyritään löytämään käytetystä datasta muuttujia, joilla on jään lisäksi merkittäviä vaikutuksia aluksen kulkemiseen jään peittämässä Itämeressä.

Tämä työ on osa AIGA-tutkimushanketta (*eng. Artificial Intelligence Governance and Auditing*), joka keskittyy tekoälyn tuottamien päätösten läpinäkyvyyden ja ymmärrettävyyden vahvistamiseen.

1.3 Työn rajaus ja tutkimuskysymykset

Tutkimuksessa käytettävä data on rajattu Pohjanlahdella kulkeviin laivoihin vuosilta 2018–2020. Lisäksi työn pääpaino on tekoälyn selitettävyyden tutkimisessa, eikä varsinaisiin tekoälymalleihin keskitytä syvällisemmin tässä tutkimuksessa. Työssä käydään kuitenkin läpi tekoälyn teoriaa, sillä sen ymmärtäminen on oleellista työn lukemisen kannalta. Tekoälyn regulointi on myös ollut puheenaiheena viime aikoina EU:n tulevien säännösten vuoksi, joten työssä pyritään tutkimaan tekoälyn selitettävyyttä myös regulaation näkökulmasta. Tutkimuksen tutkimuskysymyksiksi valikoituivat:

- Miksi tekoälyn selitettävyyttä tarvitaan talvimerenkulun tekoälyratkaisuissa?
- Miten selitettävä tekoäly auttaa koneoppimismallien kehittämisessä?
- Kuinka hyvin tekoälyselitteitä voidaan ymmärtää eri sidosryhmien näkökulmasta?

Tekoäly on tutkimusalueena hyvin laaja ja kattaa valtavasti eri asioita, joten työ rajattiin tekoälyn selitettävyyteen. Tutkimuksessa pyritään antamaan lukijalle ymmärrys siitä, miksi selitettävää tekoälyä tarvitaan talvimerenkulun tekoälyratkaisuissa sekä miten selitettävä tekoäly hyödyntää itse mallien kehittämisessä. Selitettävästä tekoälystä puhuttaessa on myös syytä huomioida eri sidosryhmät, sillä selitettävyys tarkoittaa eri rooleissa työskenteleville eri asioita. Täten tutkimuksessa pyritään vastaamaan myös kysymykseen siitä, miten eri sidosryhmät tulisi huomioida selitteitä luotaessa.

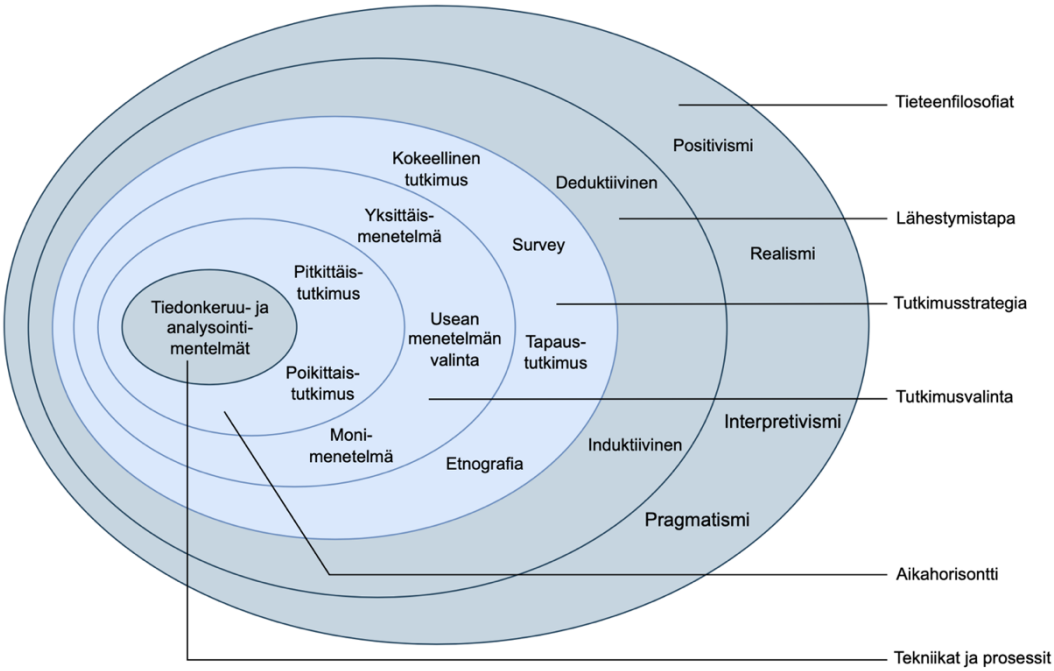
1.4 Työn rakenne

Työ rakentuu viidestä eri luvusta. Johdannossa käydään läpi tutkimuksen taustaa, rajausta ja tutkimuskysymyksiä ja viitekehystä sekä Itämeren talvimerenkulkua yleisellä tasolla. Tutkimuksen teoreettista osuutta käsitellään toisessa ja kolmannessa luvussa. Toisessa luvussa keskitytään tutkimaan tekoälyä yleisemmällä tasolla merenkulussa ja antamaan lukijalle kuva muun muassa nykyisistä tekoälyratkaisuista laivaliikenteessä

sekä sen haasteista. Luvussa myös pyritään avaamaan tekoälyä talvimerenkulun näkökulmasta, jolloin meren jäätymisellä on vaikutusta tekoälyratkaisujen tekemisessä. Kolmannessa luvussa perehdytään selitettävään tekoälyyn (XAI). Luvussa pyritään antamaan laaja yleiskäsitys siitä, mitä selitettävällä tekoälyllä tarkoitetaan ja mitä se tarkoittaa eri sidosryhmille. Tässä luvussa käsitellään myös varsinaisia tekoälyn visuaalisia selitteitä sekä pyritään tuomaan ajankohtaista näkökulmaa huomioimalla tekoälyregulaatio. Neljännessä luvussa käsitellään tutkimuksen empiiristä osiota eli miten tutkimusta on lähestytty sekä käytetyn datan keruuta ja analysointia. Viimeisessä eli viidennessä luvussa käydään läpi tutkimuksen tulokset sekä niiden arviointi ja pohditaan mahdollisia jatkotutkimusmahdollisuuksia.

1.5 Tutkimuksen viitekehys

Saunders et. al (2009) kehittämän niin sanotun sipulimallin (kuva 1) (eng. *research onion*) avulla tutkimusta voidaan tarkastella eri näkökulmista. Saundersin sipulimallia hyödynnetään siten tässäkin tutkimuksessa. Malli koostuu kuudesta eri kerroksesta, joita ovat tieteenfilosofia, lähestymistapa, tutkimusstrategia ja -valinnat, aikahorisontti sekä tiedonkeruu- ja analysointimenetelmät. Sipulimallia lähestytään uloimmasta kerroksesta sisimpään kerrokseen näiden tarkentuessa ydintä lähestyttäessä. Seuraavissa alaluvuissa käydään tarkemmin sipulimallin kerrokset tämän tutkimuksen näkökulmasta ja perustellaan miksi kyseisiin valintoihin on päädytty.



Kuva 1: Sipulimalli (Mukaillen Saunders et al. 2009)

1.5.1 Tieteenfilosofia

Saunders et al. (2009) mukaan tieteenfilosofian valinnalla on keskeinen merkitys siihen, miten koko tutkimusprosessia lähdetään viemään eteenpäin ja se luo pohjan sipulimallin sisempien kerrosten valinnoille. Tieteenfilosofiaa voidaan lähestyä kahdesta eri suuntauksesta, joita ovat ontologia sekä epistemologia. Ontologinen suuntaus käsittelee todellisuuden luonnetta ja pyrkii vastaamaan tutkijan olettamuksiin maailmasta ja kysymyksiin kuten *”Mitä tiedämme?”* sekä *”Miten tiedämme?”*. Epistemologia puolestaan pyrkii vastaamaan siihen, että mitä ja miten tietoa on saatavissa ja onko tieto pätevää eli epistemologian keskeinen painoarvo on tiedon käsitteessä. (Saunders et al. 2009; Sirén & Pekkarinen 2017) Tämä tutkimus noudattaa epistemologista suuntausta, sillä tutkimuksessa keskitytään pätevän tiedon etsimiseen ja hyödyntämiseen, eikä olettamuksiin.

Saunders et al. (2009) esittävät lisäksi neljä keskeistä tutkimussuuntausta, joita ovat positivismi, realismi, interpretivismi sekä pragmatismi. Positivismilla tarkoitetaan tiedonkeruutapaa, joka pohjautuu objektiivisuuteen ja mitattuihin arvoihin sekä havainnoituihin syy-seuraus-suhteisiin. Positivismille on myös tyypillistä, että tutkijan omat arvot vaikuttavat mahdollisimman vähän tutkimuksen tuloksiin. Positivismin suuntaus on yleisesti käytetty epistemologinen valinta. (Sirén & Pekkarinen, 2017) Realismi puolestaan pohjautuu aistihavainnoista saatuun tietoon ja painottaa todellisuutta objektiivisena käsitteenä. Havaitut ilmiöt tarjoavat uskottavaa tietoa ja suoran realismin (*eng. Direct realism*) mukaan epäsopiva tieto tarkoittaa virheitä aisteissa. Toisaalta kriittinen realismi (*eng. Critical realism*) huomioi myös aistien virheelliset tulkinnat. Interpretivismissä sosiaaliset ilmiöt tulkitaan subjektiivisiksi käsityksiksi eli ihmisen oman ajattelumallin merkitys korostuu. Pragmatismissa puolestaan korostuvat kokemus sekä käytännönläheisyys. (Saunders et al., 2009; Sirén & Pekkarinen, 2017)

Tässä tutkimuksessa hyödynnetään laajaa strukturoitua kvantitatiivista jää- ja sääolosuhteiden sekä paikkatiedon yhdistävää dataa tutkimuksen pohjana sekä perustetaan tulokset objektiivisiin näkökantoihin, joten tieteenfilosofiseksi suuntaukseksi tässä työssä on valittu positivismi. Toisaalta selitettävässä tekoälyssä voidaan nähdä myös subjektiivisia piirteitä ja Lin et al. (2020) mainitsevat, että nykyiset arviointimenetelmät vaativat subjektiivista panosta ihmisiltä eli tutkimuksessa voidaan nähdä myös pragmatismin piirteitä.

1.5.2 Tutkimuksen lähestymistapa

Sipulimallin toisena kerroksena on tutkimuksen lähestymistapa. Lähestymistavan valinta on tärkeä osa tutkimusta, sillä se kertoo päätöksistä, joita tehdään liittyen tiedon keräämiseen ja analysointiin. Saunders et al. (2009) jakavat lähestymistavat deduktiiviseen ja induktiiviseen päättelyyn. Deduktiivinen päättely on vahvasti sidoksissa luonnontieteiden kanssa, jossa erilaiset lait muodostavat pohjan selityksille ja mahdollistavat ilmiöiden ennakoinnin Deduktiivinen päättely on teorialähtöistä eli teorian pohjalta pyritään luomaan testattavia hypoteeseja (Saunders et al., 2009; Collis & Hussey, 2003). Induktiivisessa lähestymistavassa tutkimuksen pohjana ovat puolestaan aineistot ja havainnot, jolloin tavoitteena on saada ongelman luonteesta selkeämpi käsitys ja täten pyrkiä luomaan teoria sen pohjalta (Anttila, 1998; Saunders et al., 2009).

Tiivistetysti voisi siis sanoa, että deduktiivinen päättely alkaa teoriasta ja induktiivinen puolestaan kerätystä aineistosta. Tässä tutkimuksessa voidaan nähdä piirteitä molemmista lähestymistavoista, sillä tutkimus lähti liikkeelle datan tutkimisella ja siten uuden teorian luomisesta talvimerenkulussa, mutta samalla kuitenkin talvimerenkulun teorian pohjalta ja siihen perustuvien hypoteesien testauksella. Tämän vuoksi tutkimuksen lähestymistapana voidaan pitää molempia yhdistävää, niin sanottua abduktiivista päättelyä. Anttila (1998) toteaa, että abduktiivisessa päättelyä noudattavassa tutkimuksessa uuden teorian kehittäminen on mahdollista vain silloin, kun tehtyihin havaintoihin liittyy johtoajatus, joka voi olla luonteeltaan intuitiivinen käsitys tai pidemmälle muotoiltu hypoteesi.

1.5.3 Tutkimusstrategia ja -valinnat

Tutkimusstrategian valinta on Saunders et al. (2009) sipulimallin kolmas vaihe, jolla tarkoitetaan tutkimuksessa käytettävien menetelmällisten ratkaisujen kokonaisuutta, jolla tutkimusta pyritään toteuttaa (Lähdesmäki et al., 2014). Tutkimusstrategioita ovat esimerkiksi kokeellinen tutkimus, survey-tutkimus, tapaustutkimus sekä etnografia. Saaranen & Puusniekka (2006) mukaan tapaustutkimuksessa voidaan tutkia muun muassa yksittäistä tapahtumaa tai rajattua kokonaisuutta. Tapaustutkimus pyrkii myös vastaamaan kysymyksiin: miten tapahtui ja miksi tapahtui? Tässä tutkimuksessa tutkitaan niin ajallisesti, alueellisesti kuin määrällisesti rajattua laivojen kokonaisuutta sekä pyritään löytämään yksittäisiä tapahtumia data-aineistosta, joten tutkimusstrategiaksi on valittu tapaustutkimus. Tapaustutkimus tukee myös laadullisia keinoja, kuten keskusteluja, joten se on sopiva valinta tutkimukselle.

Tutkimusstrategian jälkeen valitaan tutkimukselle sopiva menetelmäsuuntaus eli käytetäänkö tutkimuksessa kvantitatiivista, kvalitatiivista vai monimenetelmällistä suuntausta. Kvantitatiivinen eli määrällinen tutkimus pohjautuu numeeriseen aineistoon ja kvalitatiivinen eli laadullinen tutkimus puolestaan ei-numeeriseen aineistoon. Saunders et al. (2009) sipulimallissa on esitetty kolme erilaista valintaa: yksittäismenetelmään, monimenetelmään sekä usean menetelmän valintaan. Yksittäismenetelmässä (*eng. mono method*) käytetään joko määrällistä tai laadullista aineistoa ja monimenetelmätutkimuksessa (*eng. mixed methods*) käytetään sekä määrällistä, että laadullista aineistoa. Jos tutkimuksessa on tarkoitus hyödyntää määrällistä tai laadullista aineistoa useampaan kertaan, puhutaan usean menetelmän valinnasta (*eng. multi-method*) (Saunders et al. 2009) Tässä tutkimuksessa käytetään kvantitatiivista aineistoa eli AIS-dataa sekä laadullista aineistoa eli haastatteluja, joten tutkimusta voidaan pitää monimenetelmätutkimuksena.

1.5.4 Aikahorisontti

Viidentenä vaiheena sipulimallissa on valita tutkimuksen aikahorisontti, jolla viitataan tutkimuksessa käytettävän aineiston keräämiseen käytettävää aikaa. Saunders et al. (2009) jakavat aikahorisontin pitkittäistutkimukseen (*eng. longitudinal*) sekä poikittaistutkimukseen (*eng. cross-sectional*). Poikittaistutkimuksessa ilmiötä tai ilmiöitä tutkitaan tiettynä ajankohtana, joten myös aineistoa kertyy vain tietyltä ajanhetkeltä tai lyhyeltä ajalta. Pitkittäistutkimuksessa nimensä mukaisesti aineistoa kerätään pidemmältä aikaväliltä ja tutkitaan ilmiön muutosta ja kehitystä.

Vaikka tutkimuksessa kerätään aineistoa pitkältä ajanjaksolta, jopa useammalta vuodelta, noudattaa tutkimus kuitenkin enemmän poikittaistutkimuksen periaatteita kuin pitkittäistutkimuksen. Toki esimerkiksi talvi- ja kesädatan muutosta tulee huomioida, sillä koneoppimismallin tulisi ennustaa erilailla talvelle kuin kesälle, mutta varsinaisia ilmiöitä, kuten laivojen pysähdyksiä, tutkitaan tiettyinä lyhyinä ajankohtina. Myös tekoälyn selitettävyyden teoreettista aineistoa kerätään lyhyeltä aikaväliltä jo pelkästään aiheen uutuuden vuoksi.

1.5.5 Tiedonkeruu ja -analysointimenetelmät

Tutkimuksessa käytettävä aineisto koostuu kirjallisuudesta, kvantitatiivisesta datasta sekä kvalitatiivisesta keskusteluaineistosta. Kirjallisuuskatsausta lähestytään Arlene

Finkin (2014) systemaattisen kirjallisuuskatsauksen mallilla. Prosessimalli koostuu seitsemästä eri vaiheesta, joita ovat tutkimuskysymysten asettiminen, tietokantojen ja sivustojen valinta, hakutermien valinta, käytännön seulan asettaminen, metodologisen seulan asettaminen, katsauksen suorittaminen sekä synteesin tekeminen tuloksista. (Salminen, 2011) Tässä työssä haetaan kirjallisuutta Tampereen yliopiston Andor-hakupalvelusta ja Google Scholar -palvelusta sekä myös Väyläviraston sivuilta. Hakutermistö pidetään englannin- ja suomenkielisenä. Kirjallisuutta käydään läpi huolellisesti ja lähteitä seulotaan tarkasti noudattaen asetettuja kriteereitä.

Kvantitatiivinen data koostuu AIS-datasta, jota on kerätty Väyläviraston tietovarastosta sekä Hakolan (2020) luomasta AIS-datasetistä. Lisäksi käytetään Suomen Ilmatieteenlaitoksen säädataa ja Ruotsin Ilmatieteenlaitoksen avointa jäädataa. Kvalitatiivista aineistoa on saatu useista talvimerenkulun asiantuntijoiden kanssa käydyistä keskusteluista. Työssä käytetään analysointimenetelmänä CRISP-DM-menetelmästä johdettua Agile CRISP-DM -menetelmää, josta kerrotaan tarkemmin analyysiluvussa.

2. TEKOÄLY JA LAIVALIIKENNE

Tässä luvussa käydään läpi laivaliikennettä Itämerellä yleisellä tasolla kuin myös talvimerenkulun näkökulmasta. Lisäksi syvennytään tekoälyyn ja koneoppimiseen sekä miten niitä voidaan hyödyntää laivaliikenteessä. Lopuksi esitetään nykyisiä talvimerenkulun tekoälyratkaisuja.

2.1 Laivaliikenne Itämerellä

Kuten johdannossa todettiin, laivaliikenne on merkittävässä roolissa Suomen ulkomaankaupassa, joten laivaliikenteen, niin kesällä kuin talvella, tulee olla turvallista ja tehokasta. Talvimerenkulussa vallitseva jäätilanne on kuitenkin merkittävässä roolissa. Toivola (2016) mainitsee turvallisen ja tehokkaan talvimerenkulun kannalta viisi keskeistä komponenttia:

1. Tarkka jäätilanneinformaatio
2. Jäärajoitukset ja rannikkokoordinaatit
3. Kauppalaivojen jääluokat ja jääkapasiteetti
4. Jäänmurtajien kapasiteetti
5. Ihmisten taidot

Talvimerenkulun asiantuntijan kanssa käydyssä keskustelussa ilmeni, että yksi suurimmista haasteista talvimerenkulussa onkin tarkka jääinformaatio, sillä jääkenttä voi muuttua jopa minuuteissa. Väyläviraston (2020) mukaan Suomen Ilmatieteenlaitos seuraa jäätilannetta päivittäin ja laatii jäätilannekarttoja sekä jäätilanteen kehitysennusteita. Väyläviraston käytössä Itämerellä työskentelee kahdeksan jäänmurtajaa: Urho, Sisu, Otso, Kontio, Voima, Fennica, Nordica sekä uusimpana Polaris. Arctian (2016) jäänmurron prosessikuvauksen mukaan jäänmurtajat ovat sijoitettu pitkin Suomen rannikkoa. Kauppa-alusten tulisi pyrkiä etenemään meriliikenneohjauskeskuksen ohjeiden avulla itsenäisesti, mutta aluksen jäädessä kiinni jäänmurtaja tekee avustuspäätöksen joko suoraan tai meriliikenneohjauskeskuksen kautta. Alukset saavat avustuksen siinä järjestyksessä, jossa jäänmurtajat ovat saaneet tiedon tai mahdollisesti jäänmurtajan päällikön parhaaksi katsomassa järjestyksessä kokonaislogistiikan kannalta. Jäänmurtaja aloittaa vähentämällä jääkentän puristusta tekemällä rännin kauppa-aluksen sivulle, jonka jälkeen kauppa-alus voi pyrkiä eteenpäin. Päästyään liikkeelle avustettavan aluksen on käytettävä konetehoa jäänmurtajan ohjeiden mukaisesti - usein kuitenkin

täydellä teholla. Tuuli tuo lisähaasteen avustukseen, sillä tuulen vaikutuksesta ränni voi sulkeutua uudelleen ja siten aiheuttaa jääkenttään puristusta, jolloin kauppa-alus voi jäädä uudelleen kiinni jäähän, vaikka se seuraisikin jäänmurtajaa. Aluksen jäädessä uudestaan jumiin, voidaan myös aloittaa hinaus, jolloin jäänmurtaja antaa myös veto-apua avustettavalle alukselle. Usein avustettavia aluksia on useampia, jolloin kyseessä on saattue tai niin sanottu konvoi, jolloin kunkin aluksen on pidettävä turvallista etäisyyttä toisiinsa.

Jotta laivaliikenne talvella olisi mahdollisimman tehokasta ja turvallista sekä aluksilla kyky liikkua jäissä, talvimerenkulussa on käytössä jääluokitus aluksille. Aluksen jääluokkaan vaikuttavat muun muassa aluksen koneteho sekä erilaiset kuljetuskoneiston ja rungon vahvistukset. Jääluokituksella on vaikutusta niin väylämaksun suuruuteen kuin oikeuteen saada jäänmurtoavustusta. Suomessa on käytössä seuraavat taulu 1 esitetyt jääluokat (Traficom, 2022):

Jääluokka	**Kulkeminen jäissä**
IA Super	Vaikeat jääolosuhteet ilman avustusta
IA	Vaikeat jääolosuhteet avustettuna
IB	Keskivaikeat jääolosuhteet
IC	Helpot olosuhteet
II	Ei täytä jääluokkamääräysten vaatimuksia.
III	Ei täytä jääluokkamääräysten vaatimuksia.

Taulu 1: Alusten jääluokat (Traficom, 2022; Väylävirasto, 2020)

IA Super -jääluokan alukset pystyvät kulkemaan vaikeissa jääolosuhteissa pääosin ilman avustusta ja IA jääluokkien alukset pystyvät kulkemaan vaikeissa jääolosuhteissa tarpeen mukaan avustettuna. IB-luokan alukset puolestaan keskivaikeissa ja IC-luokan alukset helpoissa jääolosuhteissa. II- sekä III-jääluokan alukset eivät täytä jääluokkamäärityksen vaatimuksia ja ovat tarkoitettu Itämerta helpompiin jääolosuhteisiin. II-jääluokan alukset voivat saada jäänmurtoavustusta, jos niiden kantavuus on riittävä ja jääolosuhde riittävän helppo, mutta III-luokan alukset eivät voi milloinkaan saada avustusta. Avustettavat alukset ovat kuitenkin itse yksin vastuussa omasta navigoinnistaan sekä myös avun ja neuvon vastaanottaminen on aluksen omalla vastuulla. Avovesikauteen verrattuna talvimerenkulussa alukset ovat siis alttiimpia isommille riskeille muun

muassa saattueiden ja rikkoutuneen jääkentän sekä näistä johtuvan haastavan etäisyydenhallinnan vuoksi, joka lisää yhteentörmäyksien vaaraa. Tämän vuoksi jääluokituksia ja muita sääntöjä on hyvin tärkeä noudattaa. (Traficom, 2022; Väylävirasto, 2020)

2.2 Mitä on tekoäly?

Tekoäly on noussut vuosi vuodelta yhä enemmän osaksi elämäämme ja jokapäiväistä keskustelua, vaikkakin tekoälyä onkin tutkittu jo 1940-luvulta lähtien. Varsinaisen tekoälyn termin kuitenkin kehitti John MacCarthy vuonna 1956. (Pietikäinen & Sivén, 2021; Taulli, 2019). Tekoälyn määritelmää voidaan lähestyä monista eri näkökulmista ja erilaisia määritelmiä tekoälylle voidaankin sanoa olevan yhtä monta kuin on kirjoittajia. Russel & Norvig (2016) esittävät kahdeksan käytössä olevaa eri määritelmää (Taulu 2) kahden dimension suhteen. Taulukon yläosassa esitetyt määritelmät koskevat ajatteluprosesseja ja päättelyä, kun taas näiden alla esitetyt määritelmät koskevat käyttäytymistä. Taulukon vasemman puoliset määritelmät puolestaan mittaavat tarkkuutta verrattuna ihmisten ajatteluun ja oikeanpuoleiset määritelmät vastaavat ihanteellista suorituskykymittausta ja rationaalisuutta. Historiallisesti kaikkia neljää tekoälyn lähestymistapaa ollaan Russel & Norvigin (2016) mukaan käytetty. Ihmiskeskeisen lähestymistavan on oltava osa empiiristä tiedettä, joka sisältää hypoteeseja ja havaintoja ihmisten käyttäytymisestä ja rationaalisen lähestymistavan sisällettävä yhdistelmän matematiikkaa ja tekniikkaa.

Eräs tunnettu inhimillisen toiminnan lähestymistapa on niin sanottu Turingin (1950) testi, jossa testataan ihmistarkkailijan kykyä erottaa tietokoneen ja ihmisen antamat vastaukset. Jos koneen vastauksia ei pystytä erottamaan ihmisen vastauksista, kone läpäisee Turingin testin. Jotta kone voi läpäistä testin, sillä tulee olla monia eri ominaisuuksia. Muun muassa luonnollisen kielen käsittelykyvykkyyttä (*eng. Natural Language Processing, NLP*) onnistuneeseen kommunikaatioon valitulla kielellä, tiedon kuvailumenetelmiä (*eng. Knowledge presentation*) tiedon ja kuullun tallentamiseen, automaattista päättelyä (*eng. Automated reasoning*) tallennetun informaation käyttämistä kysymysten vastaamiseen sekä uusien johtopäätösten luomiseen ja koneoppimista (*eng. Machine learning*) havaitsemaan sekä ekstrapoloimaan malleja. Edellä mainittujen lisäksi myös konenäkö ja robotiikka ovat osa tekoälyä ja niin sanottua täydennettyä Turingin testiä. Nämä kuusi tieteenalaa muodostavat suurimman osan tekoälystä ja Turingin testi onkin ajankohtainen vielä yli 60 vuotta myöhemmin tänäkin päivänä. (Russel & Norvig, 2016)

Thinking Humanly	**Thinking Rationally**
“The exciting new effort to make computers think… machines with minds, in the full and literal sense.” (Haugeland, 1985) “(The automation of) activities that we associate with human thinking, activities such as decision-making, problem solving, learning…” (Bellman, 1978)	“The study of mental faculties through the use of computational models.” (Charniak and McDermott, 1985) “The study of the computations that make it possible to perceive, reason, and act.” (Winston, 1992)
Acting Humanly	**Acting Rationally**
“The art of creating machines that perform functions that require intelligence when performed by people.” (Kurzweil, 1990) “The study of how to make computers do things at which, at the moment, people are better.” (Rich and Knight, 1991)	“Computational Intelligence is the study of the design of intelligent agents.” (Poole et al., 1998) “AI ...is concerned with intelligent behavior in artifacts.” (Nilsson, 1998)

Taulu 2: Russel & Norvig (2016) esittämät tekoälyn määritelmät

Inhimillisen ajattelun perustana on tietää, että miten ihminen ajattelee. Jos sanomme, että tekoälyohjelma ajattelee kuin ihminen, on oltava jokin tapa määrittää, miten ihminen ajattelee ja siten ymmärtää ihmismielen toimintaa. Russel & Norvig (2016) esittävät kolme keinoa, joita ovat itsetutkiskelu, psykologiset kokeet sekä aivokuvaus. Kun on olemassa tarpeeksi tarkka mielen teoria, se on mahdollista esittää myös tietokoneohjelmana. Jos ohjelman laitetun syötteen ja saadun tuotoksen käyttäytyminen vastaa ihmisen käyttäytymistä, se on todiste, että jotkin ohjelman mekanismit voivat toimia myös ihmisissä. Tekoälyn alkuaikoina lähestymistapojen välillä olikin usein sekaannusta. Saatettiin väittää, että algoritmi suorittaa tehtävänsä hyvin ja on siten verrattavissa ihmiseen. Nykypäivänä kuitenkin nämä erotetaan toisistaan, joka on mahdollistanut tekoälyn, että kognitiivisen tieteen kehittymisen nopeammin. (Russel & Norvig, 2016)

Inhimillisen ajattelun ja toiminnan lisäksi Russel & Norvig (2016) kirjoittavat rationaalisesta ajattelusta ja toiminnasta. Rationaalisen ajattelun tavoitteena on perustaa ohjelmat ”oikealle ajattelulle”. Tällä lähestymistavalla on kuitenkin kaksi pääasiallista estettä. Kaikkea tietoa ei voida esittää loogisilla merkeillä, varsinkaan, kun tieto ei ole täysin varmasti oikeaa ja toiseksi, on suuri ero ratkaistaanko ongelmaa periaatteen tasolla vai

käytännön tasolla. Jo muutaman sadan faktan ongelmat (parempi suomennos) voivat viedä jokaisen tietokoneen laskentaresurssit, jos ei ole olemassa ohjeita siitä, minkälaista päättelyvaiheita tulisi kokeilla ensin. (Katso vielä vika lause). Rationaalista toimintaa Russel & Norvig (2016) lähestyvät niin sanotun rationaalisen agentin kautta. (rationaalisesta agentista löytyy lähteitä). Tässä merkityksessä agentilla tarkoitetaan yksinkertaisesti vain jotain, joka tekee jotakin, kuten toimii itsenäisesti, havaitsevat ympäristöänsä sekä soputuvat muutoksiin ja luovat tavoitteita. Rationaalisella agentilla tarkoitetaan puolestaan agenttia, joka toimii parhaan lopputuloksen saavuttamiseksi tai epävarmassa tilanteessa parhaan odotetun lopputuloksen saavuttamiseksi.

Rationaalisessa ajattelussa ajattelun lakeihin (*eng. laws of thought)* pohjautuvassa lähestymistavassa tekoälyssä korostetaan oikeita johtopäätöksiä. Oikeiden johtopäätösten tekeminen on joskus myös osa rationaalisen agentin toimintaa, koska rationaaliseen toimintatapaan kuuluu kyky loogisesti perustella, että tietyn toiminnon suorittaminen johtaa tietyn tuloksen saavuttamiseen ja siten toteuttaa tämä toiminto. Toisaalta oikea päättely ei ole pelkästään rationaalisuutta; joissain tilanteissa ei ole todistetusti oikeaa tapaa toimia, mutta jotain on silti tehtävä. On myös tapoja toimia rationaalisesti ilman, että se sisältää harkintaa ja päättelyä. Rationaalisen agentin lähestymistavalla on kaksi hyötyä verrattuna muihin lähestymistapoihin. Ensinnäkin se on yleisempi kuin ajattelun lakeihin perustuva lähestymistapa, koska oikea päättely on vain yksi useista mahdollisista keinoista rationaalisuuden saavuttamiseksi. Toiseksi se soveltuu paremmin tieteelliseen kehitykseen kuin ihmisten käyttäytymiseen ja ajatteluun pohjautuvat lähestymistavat. (Russel & Norvig, 2016)

Tekoäly voidaan jakaa myös niin sanottuun heikkoon ja vahvaan tekoälyyn. Siukonen ja Neitaanmäki (2019) mukaan heikko tekoälyn toiminta perustuu älykkäisiin ja taitavasti toimiviin algoritmeihin, jolloin koneoppimiseen perustuvat tietokoneohjelmistot kykenevät toimivaan ilman ymmärrystä tehtävästä asiasta. IBM:n (2020) mukaan esimerkiksi shakkiohjelma, joka perustaa siirtonsa valmiisiin käskyihin on heikkoa tekoälyä, sillä se kykene arvioimaan ja muokkaamaan siirtojansa itsenäisesti pelin edetessä, vaan tekee siirrot valmiin algoritmin mukaisesti. Jos shakkiohjelma kykenisi opettamaan itse itseänsä ja kykenisi toimimaan ihmisen tavoin, olisi kyseessä vahva tekoäly. Vahvan tekoälyn tavoitteena on luoda älykkäitä koneita, joita ei voi erottaa ihmismielestä. Tällaisen tekoälykoneen kehittäminen on verrattavissa myös ihmiseen, sillä sen olisi opittava syötteen ja kokemusten kautta, kehittyen jatkuvasti ja parantaen kykyjään ajan myötä. Vaikka tutkijat sekä akateemisella että yksityisellä sektorilla panostavat vahvan tekoälyn kehittämiseen, se on kuitenkin vielä teoreettinen konsepti konkreettisen todellisuuden sijaan.

On myös käyty keskustelua siitä, että onko vahvaa tekoälyä mahdollista edes luoda ennen kuin menestyksen mittareita, kuten älykkyyttä ja ymmärrystä on tarkasti määritelty. Siukonen ja Neitaanmäki (2019) jatkavat, että tutkijoiden tulisi pystyä luomaan jotain täysin ennennäkemätöntä sekä löytää vastaukset erilaisiin kysymyksiin, kuten ”*voiko kone olla tietoinen tilastansa”* sekä ”*voiko kone tuntea reaalimaailman ja saavuttaa olotilan, jossa se määrittää omat pyrkimyksensä ja tavoitteensa”*? Vasta kun tällaiset kysymykset saadaan muutettua matematiikan lainalaisuuksia noudattaen tekoälyn ymmärtämään muotoon, on tuloksia mahdollista saada.

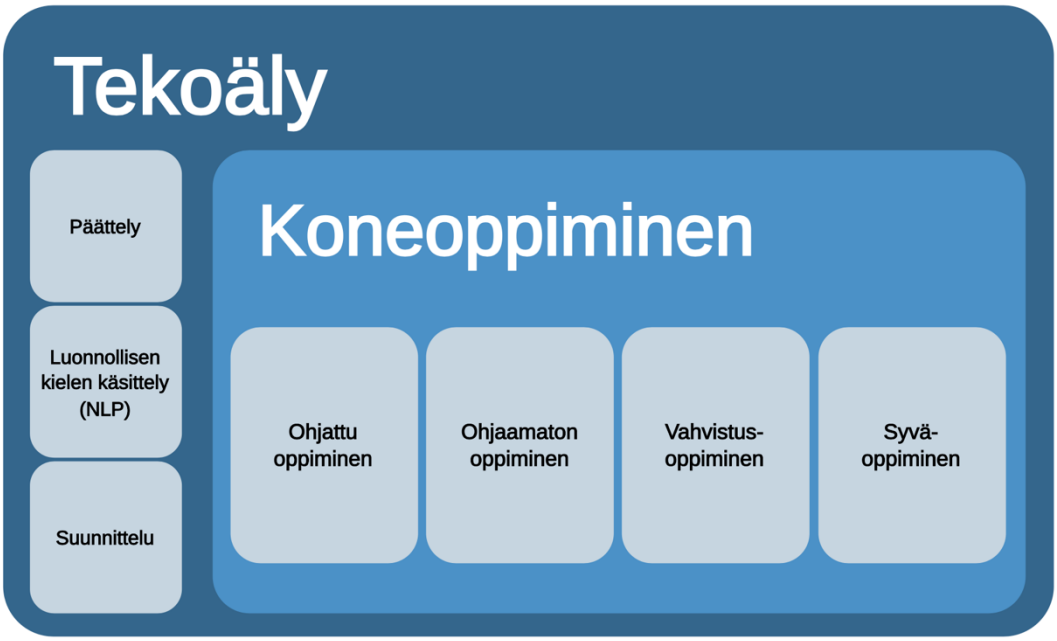
Liu (2021) mukaan heikon ja vahvan tekoälyn suhde on metaforisesti kuin lentävien koneiden ja lintujen vertailu. Linnut toimivat paljon kehittyneemmin kuin nykyajan edistynein lentävä kone, sillä ne voivat joustavasti muuttaa käyttäytymistään, kun taas lentävät koneet eivät. Ihmiset saattavat ajatella, että on epätodennäköistä eikä välttämätöntä, että lentävät koneet kehittyisivät lintujen kaltaisiksi, jolloin samalla periaatteella voidaan myös ajatella, että heikon tekoälyn on epätodennäköistä eikä välttämätöntä kasvaa vahvaksi tekoälyksi. Liu jatkaa, että heikon tekoälyn suurin arvo piilee siinä, että se tarjoaa skaalautuvia, vähemmän työtä vaativia, tarkkoja sekä yleistettäviä työkaluja. Vaikka heikolla tekoälyllä ei ole todellista älykkyyttä, se täyttää suurelta osin datan käsittelyyn liittyvät tarpeet.

2.3 Koneoppiminen

Koneoppimisella tarkoitetaan tekoälystä ja tietojenkäsittelytieteestä haarautuvaa alaa, joka pyrkii datan ja algoritmien käytöllä jäljittelemään tapaa, jolla ihmiset oppivat. Koneoppimisessa käytetään erilaisia algoritmeja, jotka oppivat iteratiivisesti datasta parantaakseen, kuvatakseen dataa sekä ennustaakseen tuloksia. Kun algoritmeille syötetään opetusdataa, algoritmien on mahdollista tuottaa tarkempia malleja. Koneoppimismalli on tulos, joka syntyy, kun algoritmia opetetaan opetusdatalla. Tämän jälkeen, kun koneoppimismallille antaa syötteen, se antaa tulosteen. Esimerkiksi ennustava algoritmi luo ennakoivan mallin ja kun mallille syöttää dataa, saa tulosteeksi malliin syötettyyn dataan perustuvan ennusteen. (Hurwitz & Kirsch, 2018)

Shalew-Shwartz & Ben-David (2014) mainitsevat kaksi eri näkökulmaa, jolloin koneoppimista tarvitaan sen sijaan, että tietokone ohjelmoitaisiin tekemään haluttu tehtävä. Näitä ovat tehtävät, jotka ovat liian monimutkaisia ohjelmoitaviksi sekä tarve mukautumiselle. Liian monimutkaisia ovat esimerkiksi tehtävät, joita ihmiset tekevät rutiininomaisesti, kuten ajaminen, puheentunnistus ja kuvien tunnistus. Kaikissa näissä tehtävissä huipputason koneoppimismallit saavuttavat jo varsin tarkkoja tuloksia, kun ne al-

tistetaan riittävälle määrälle opetusdataa. Lisäksi tehtäviin, jotka ylittävät ihmisten kyvyn, voi koneoppiminen tarjota lisähyötyä. Tällaisia ovat muun muassa erittäin suurten ja monimutkaisten datakokoelmien analysointi, kuten esimerkiksi tähtitieteeseen, lääketieteeseen ja hakukoneisiin liittyvä massadata, jotka ovat ihmisille mahdottomia ymmärtää. Toinen näkökulma on tarve mukautuvuudelle. Yksi ohjelmoitujen työkalujen rajoittava ominaisuus on niiden jäykkyys – kun ohjelma on koodattu ja asennettu, se pysyy ennallaan. Monet tehtävät kuitenkin muuttuvat ajan myötä tai niiden käyttäjät vaihtuvat.



Kuva 2: Tekoälyn kokonaisuus (Mukaillen Hurwitz & Kirsch, 2018)

Koneoppimiseen perustuvat ohjelmat, jotka mukautuvat annetun syötteen perusteella, tarjoavat keinoja tällaisten ongelmien ratkaisemiseen. Esimerkiksi sovellus, joka tunnistaa roskaposteja, mutta joka myös kykenee mukautumaan roskapostin luonteen muutoksiin. (Shalew-Shwartz & Ben-David, 2014)

Koneoppimisen yhteydessä puhutaan usein muun muassa tekoälystä ja syväoppimisesta. Kuvasta 2 nähdään, että koneoppiminen on osa isompaa tekoälyn kokonaisuutta, joka voidaan ymmärtää laajimpana tapana kuvata ajatteluun kykeneviä järjestelmiä. Koneoppiminen on yksi neljästä tekoälyn osajoukosta ja itse koneoppimisen alle kuuluvia joukkoja ovat ohjattu oppiminen (*eng. supervised learning*), ohjaamaton oppinen (*eng. unsupervised learning*), vahvistusoppiminen (*eng. reinforcement learning*) sekä syväoppiminen (eng*. deep learning*) sekä siihen liittyvät neuroverkot. (Hurwitz & Kirsch, 2018)

Ohjatulla oppimisella tarkoitetaan oppimista, jossa käytetään sellaisia datasettejä algoritmien opettamiseen, joissa muuttujien arvot ovat tiedossa (*eng. labeled data*). Ohjatun

oppimisen tarkoituksena on löytää datasta malleja, joita voidaan hyödyntää analytiikkaprosessissa. Esimerkiksi datasetti voi sisältää miljoonia eläinkuvia, jolloin ohjatussa oppimisessa on keskeistä, että datasetissä myös kerrotaan mikä eläin kuvassa on. Näin ollen algoritmi kykenee tunnistamaan toistuvia piirteitä jokaisesta eläimestä ja luomaan koneoppimismallin, joka pystyy siten kertomaan mikä eläin on kyseessä. Tällaisessa kuvantunnistuksessa on kyse luokittelusta (*eng. classification*), mutta myös regressiomallit ovat hyvin suosittuja. Näitä käytetään, kun muuttujien arvot ovat jatkuvia, jolloin datasta voidaan luoda ennustemalleja. (Hurwitz & Kirsch, 2018) Lineaarinen regressioanalyysi on esimerkiksi paljon käytetty ja yksinkertainen toteuttaa. Ohjaamattomassa oppimisessa on kyse puolestaan siitä, että dataa ei ole ”otsikoitu” (*eng. unlabeled data*). Edelliseen esimerkkiin viitaten, malli ei siis voi mitenkään tietää, mikä eläin on kyseessä, sillä opetusdatassa ei ole tietoa mahdollisista eläimistä. Sen sijaan, malli voi jakaa datan eri joukkoihin, jolloin puhutaan klusteroinnista, joka on ohjaamattoman oppimisen yleisimpiä metodeja.

Vahvistusoppiminen on käyttäytymiseen perustuva oppimismalli, jossa algoritmi saa palautetta suorituksen aikana ja oppii siitä. Vahvistusoppiminen eroaa muista ohjatun oppimisen menetelmistä siten, ettei sitä kouluteta harjoitusdatalla, vaan malli oppii niin sanotusti yrityksen ja erehdyksen kautta. Siten sarja onnistuneita päätöksiä johtaa prosessin ”vahvistukseen”, koska se ratkaisee ongelman parhaiten. Vahvistusoppimista käytetään paljon muuan muassa robotiikassa. Esimerkiksi robottien opettamisessa kulkemaan rappusia dataa säädetään onnistumisten ja epäonnistumisten myötä niin kauan, että robotti kykenee kulkemaan rappusia. Syväoppiminen on erityinen koneoppimismenetelmä, joka hyödyntää neuroverkkoja oppiakseen datasta iteratiivisesti kerroksittain. Erityisesti mallien luomiseen strukturoimattomasta datasta neuroverkot ovat hyödyllisiä. Syväoppiminen on suunniteltu jäljittelemään ihmisaivojen toimintaa, jotta tietokoneet voidaan opettaa käsittelemään abstraktioita sekä huonosti määriteltyjä ongelmia. Neuroverkkoja ja syväoppimista käytetäänkin usein kuvan- ja puheentunnistuksessa sekä konenäkösovelluksissa. Vaikka syväoppiminen on hyvin samanlaista kuin perinteiset neuroverkot, käytetään syväoppimisessa enemmän piilotettuja kerroksia. Mitä monimutkaisempi ongelma, sitä enemmän piilotettuja kerroksia mallissa on. (Hurwitz & Kirsch, 2018)

Kuten edellä on mainittu, erilaiset algoritmit ovat keskeinen osa koneoppimista ja regressio- sekä luokittelualgoritmit ovat yleisesti käytettyjä. Sarker (2021) mainitsee suosituimmiksi luokittelualgoritmeiksi muun muassa Naive Bayes- (NB), Linear Discriminant Analysis- (LDA), Logistic Regression- (LR), K-nearest neighbors- (KNN), Support Vector Machines- (SVM), Decision tree- (DT) sekä Random forest- (RF) algoritmit.

Sarker (2021) listaa suosituimmiksi regressioalgoritmeiksi muun muassa lineaarisen regression ja polynomisen regression. Hänen mukaansa regressioanalyysi sisältää useita koneoppimismenetelmiä, joiden avulla voidaan ennustaa jatkuva tulosmuuttuja yhden tai useamman ennustajamuuttujan perusteella. Merkittävin ero luokittelun ja regression välillä onkin se, että luokittelija ennustaa erilaiset erilliset luokat, kun taas regressiomalli perustuu jatkuvan muuttujan ennustamiseen. Lineaarista regressiota voidaan pitää yhtenä suosituimmista koneoppimisalgoritmeista. Siinä ennustettava muuttuja on jatkuva ja riippumattomat muuttujat voivat olla joko jatkuvia tai diskreettejä ja regressioviivan muoto on lineaarinen. Polynominen regressio eroaa lineaarisesta regressiosta siten, että riippumattoman muuttujan ja ennustettavan muuttujan suhde ei ole lineaarinen. (Sarker, 2021)

Monet algoritmit, kuten esimerkiksi Random forest, toimivat niin regressioanalyysissa kuin myös luokittelussa. Tässä työssä käytettävän selitteen pohjana käytetään Random forest -regressioalgoritmiin perustuvaa koneoppimismallia laivojen nopeuden ennustamiseen. Random forest on päätöspuihin pohjautuva ohjatun oppimisen malli. Päätöspuu (*eng. Decision tree*) on malli, joka muistuttaa normaalia hierarkkista päätösprosessia. Useimmiten päätöspuissa käytetään niin sanottua binääristä päätöksentekoa, jolloin jokaisessa päätöksessä on vain kaksi vaihtoehtoa, kuten onko arvo alle vai yli tietyn raja-arvon (*eng. threshold*). Tällainen lähestymistapa on usein yksinkertaisin sekä mallin opettamisen kannalta intuitiivisin, koska jokainen päätös vähentää jopa puolella muita mahdollisia tapauksia. Lisäksi päätöspuiden etuna on niiden tulkittavuus, sillä puut voidaan muuttaa niin sanotuiksi ”jos-niin”-säännöiksi, jotka ovat helposti ymmärrettäviä. Tästä syystä päätöspuut ovat suosittuja ja käytetympiä kuin menetelmät, jotka ovat tarkempia, mutta vaikeammin tulkittavia. (Alpaydin, 2014; Bonaccorso, 2018)

Hastie et al. (2009) mukaan päätöspuiden epätarkkuus on kuitenkin yksi aspekti, joka estää niitä olemasta ideaali valinta ennustamiseen. Päätöspuut harvoin kykenevät tarkkuuteen, mitä laadukas data mahdollistaisi. Erilaisilla tehostamiskeinoilla, kuten AdaBoostilla, on mahdollista parantaa päätöspuiden tarkkuutta, mutta se vaikuttaa muun muassa mallin nopeuteen sekä tulkittavuuteen. Random forest -algoritmi on usein tarkempi vaihtoehto päätöspuille, sillä Random forest:in pohjana toimivat päätöspuut. Perusajatuksena on siis yhdistää satoja tai jopa tuhansia yksittäisiä päätöspuita yhteen malliin. Yksittäiset päätöspuut eivät siis ole välttämättä kovin tarkkoja, mutta yhdistelemällä useita satunnaisia päätöspuita päästään lähemmäksi keskiarvotulosta. Yksi keskeinen Random forest -algoritmiin liittyvä termi onkin Bootstrap-aggregointi, jolla tarkoitetaan tekniikkaa, jossa keskiarvoitetaan useampi malli varianssin pienentämiseksi ja

ylisovittamisen (*eng. overfitting*) riskin vähentämiseksi. Bootstrap-aggregointi toimii erityisen hyvin päätöspuihin, sillä ne ovat yleisesti ottaen ”säröisiä” (*eng. noisy*) Lisäksi bootstrap-aggregoinnissa jokainen luotu puu on jakautunut identtisesti, jolloin jokaisen n määrän päätöspuiden keskiarvon odotusarvo on sama kuin minkä tahansa yksittäisen puun. Tämä tarkoittaa, että bootstrap-aggregoitujen puiden harha (*eng. bias*) on siten sama kuin yksittäisten bootstrap-puiden. (Hastie et al., 2009)

2.4 Tekoäly laivaliikenteessä

Teknologian hyödyntäminen laivaliikenteessä mahdollistaa entistä parempaa ennakointikykyä sekä toiminnan tehostamista. Tällaiset toteutukset ovat sisältäneet muun muassa reaaliaikaista analytiikkaa, parannettua aikataulusta sekä automatisoituja prosesseja (DeChant, 2019). Leong (2019) haastatteleman tekoälyasiantuntija Terry Singhin mukaan merenkulkuala on ollut aina hidas uuden ja nousevan teknologian käyttöönotossa, mutta tekoäly on voi olla mahdollisuus, joka tulisi huomioida vakavasti. Meriliikenteen toiminnot ovat nykyään optimoitu hyvin tehokkaasti, mutta Singhin mukaan toimintoja on vielä mahdollista tehostaa. Tällaisia ovat muun muassa erilaiset tarkkuusoperaatiot, joissa hyödynnetään esimerkiksi maantieteellistä dataa tarkkojen saapumis- ja lähtöaikojen ennustamiseen, laivojen reitityksen ja uudelleenreitityksen hienosäätöön tai polttoaineen kulutuksen mallintamiseen. Lisäksi myös turvallisuus on alue, jossa onnettomuuksia ja tapaturmia voidaan vähentää koneoppimista hyödyntämällä.

DeChant (2019) mainitsee viisi esimerkkiä tekoälyn eduista laivaliikenteessä, joita ovat kehittynyt analytiikka, automatisointi, parantunut turvallisuus, reittioptimointi sekä suorituskyvyn ennustaminen. Kehittyneen analytiikan avulla eri tietolähteistä on mahdollista saada arvokasta liiketoimintatietoa, joka auttaa varmistamaan, että päätökset perustuvat dataan ja datalla todistettuihin menetelmiin. Automaatiolla on merkityksensä merenkulussa ja koneoppimiskyvykkyydet auttavat muun muassa historiadatan prosessoinnissa erilaisten säännöllisten mallien löytämiseksi ja tunnistamaan ongelmat ennen kuin ne tapahtuvat. Tekoälyn avulla voidaan vähentää onnettomuuksia sekä tunnistaa uhkia ja muuta epäilyttävää toimintaa. Reittien optimoinnissa voidaan rakentaa tekoälymalleja tehokkaimman reitin määrittämiseksi polttoaineenkulutus ja sääolosuhteet huomioiden sekä määrittää paras kurssi sopivalla nopeudella. Suorituskyvyn ennustamisessa voidaan esimerkiksi tutkia nopeuden ja tehon välistä suhdetta vedenalaisen lian aiheuttaman suorituskyvyn muutoksen ennustamiseksi, jolloin historiadataa voidaan hyödyntää ymmärtämään, kuinka nopeasi laivojen suorituskyky heikkenee.

Tekoälyn hyödyntäminen on myös mahdollistamassa alusten autonomisen kulkemisen. Sivadas, A & Samuel, A. (2019) mukaan on jo kehitetty autonomisia aluksia, jotka kykenevät kulkemaan satamasta satamaan ja omaan terminaaliinsa ilman ihmisen ohjausta. MacKinnon et al. (2020) lisäävät, että automaatio, joka on luotu päätöksenteon tueksi, mahdollistaa positiivista kehitystä niin työmäärässä, törmäysriskeissä kuin yleisessä turvallisuudessa. Ongelmana kuitenkin usein on, että tällaisia teknologioita otetaan käyttöön ilman, että kiinnitetään tarpeeksi huomiota käyttäjien kouluttamiseen, tekoälyn luotettavuuteen ja turvallisuuteen sekä teknologian valmiuteen. MacKinnon et al. (2020) mainitsevat, että merenkulkualalla automaation määrän nousua varten on varauduttava koneoppimisen lisäksi myös teknologian standardisointiin sekä regulaatioihin. Koska tekoäly- ja automaatioratkaisut tulevat kehittymään eri teknologian tarjoajilta, niiden standardisoinnissa ja reguloinnissa sekä käytössä navigoinnissa tulee todennäköisesti olemaan valvontaa. Teknisten kysymysten lisäksi myös filosofisia ja eettisiä asioita pitäisi käsitellä. Voivatko esimerkiksi koneet saavuttaa ihmisen älykkyyden tason, joka riittää täysin miehittämättömiin, autonomisiin järjestelmiin? Autonomiset järjestelmät ovat kuitenkin vain yksi tekoälyn osa-alue laivaliikenteessä ja erilaisia tekoälyn hyödyntämistapoja on lopulta monia.

2.5 Nykyisiä talvimerenkulun ennustemalleja

Talvimerenkulussa jää on keskeinen muuttuja, joka erottaa sen normaalista merenkulusta ja luo haasteita tekoälyratkaisujen kehittämiseen. Prithvi et al. (2021) mukaan arktisilla alueilla esimerkiksi nopeuden ennustaminen on edelleen haasteellista, kun otetaan huomioon nopeasti muuttuvat jäätilanteet ja talvimerenkulun erityispiirteet. Esimerkiksi konvoissa johtavan jäänmurtajan kapteeni määrittää nopeuden, alusten välisen etäisyyden tai moottoritilan riippuen jäätilanteesta ja alusten jääluokasta. Tämän vuoksi nopeuden ennustamisen ongelma voi liittyä yksittäisen aluksen sijaan useamman aluksen kokonaisuuteen.

Tekoälyn hyödyntämisestä talvimerenkulussa on nykyään olemassa tutkimuksia ja erityisesti nopeuden ennustamiseen jääolosuhteissa liittyvää tutkimusta on tehty. Lehtola et al. (2019) mukaan aluksen nopeutta ja suorituskykyä jääolosuhteissa voidaan mallintaa joko semiempiiristen tai dataan perustuvien mallien avulla. Semiempiiriset mallit perustuvat jään eri vastustuskomponenttien analyyttisen jakautumiseen ja simulaatioiden avulla mallinnettuun jääkuormien vaikutukseen aluksen runkoon, jolloin jään aiheuttama kokonaisvastus voidaan määrittää ja yhdistettynä työntövoimaan on mahdollista arvioida aluksen nopeus. Dataan perustuvat mallit eivät huomioi alusten ja jäiden fyysi-

siä vuorovaikutuksia, vaan pyrkivät lähinnä jäljittelemään riippuvuuksia aluksen nopeuden ja/tai todennäköisyyden välillä, että alus jää kiinni jäähän eikä kykene liikkumaan. Tätä varten käytetään yleensä AIS-datasta saatavia nopeustietoja sekä alusten mittauslaitteista tai jääennusteista saatavia jäätietoja. Montewka et al. (2014) mukaan semiempiiristen mallien ongelma on se, etteivät ne huomioi jääolosuhteiden yhteisvaikutusta laivan nopeuteen sekä jättävät pois jäänpuristuksen vaikutuksen. Jäänpuristus yhdessä jäävallien kanssa voi rajoittaa merkittävästi alusten liikkumiskykyjä ja jopa pysäyttää, vaikka jääolosuhteet olisivat muuten aluksen jääluokan rajoissa.

Montewka et al. (2014) ovat lähestyneet nopeuden ja jäähän juuttumisen ennustamista käyttämällä dataperusteisia malleja, jotka huomioivat jään ominaisuuksien, kuten tasaisen jään paksuuden ja konsentraation, jäävallien sekä jäälauttojen, yhteisvaikutuksen. Mallien luomiseen käytettiin AIS-dataa sekä Suomen Ilmatieteenlaitoksen HELMI-jääennustemallia ja laivan suorituskyvyn sekä jääolosuhteiden välisen suhteen määrittelyyn Bayes-verkkoja. Mallien avulla päästiin keskimäärin 80% tarkkuuteen aluksen nopeuden ennustamisessa sekä keskimäärin 90% tarkkuuteen jäähän jumittumisen ennustamisessa. Kyseiset mallit ovat kuitenkin päteviä vain tietynlaiselle alukselle ja käytetylle HELMI-jäämallille sekä tuloksia voidaan tulkita vain mallien luomiseen käytetyn jääolosuhteiden rajoissa, joka tässä kuvastaa pohjoisen Itämeren ankaria jääolosuhteita.

Hakola (2020) on työssänsä tutkinut ja tehnyt talvimerenkulun mallinnusta lähestyen sitä meren, aluksen kulkemisen, reitin ja nopeuden mallintamisella sekä kehittänyt mallit reitin ja meriliikenteen ennustamiselle. Meren mallinnus toimii pohjana alusten liikkumisen mallintamiselle, jolloin valittu merialue on jaettu yhtä suuriin, 2,5km x 2,5km tai 5km x 5km kokoisiin alueisiin riippuen halutusta tarkkuudesta. Reitin mallinnuksessa pyritään löytämään aluksen tyypillisiä liikkumismalleja, jolloin reitin mallinnus tapahtuu laskemalla todennäköisyys siirtyä toiselta alueelta toiselle. Tämän jälkeen on mahdollista mallintaa nopeus, jota käytetään aluksen paikan sekä reitin ennustamisessa. Nopeuden mallintamisessa lasketaan aluksen keskiarvonopeus jokaisessa solmukohdassa. Mallintamisten jälkeen voitiin tehdä reitin ja liikenteen ennustaminen. Nopeuden ennustamiseen käytettiin A*-algoritmia, joka on laskennallisesti tehokkaampi verrattuna muihin lyhyimmän polun algoritmeihin. Normaaliin, ilman jäitä tapahtuvaan merenkulkuun verrattuna talvimerenkulussa reitin ennustaminen on selkeästi tarkempaa, sillä jäänmurtajien määrittämät reitit (*eng. dirway*) määrittävät hyvin pitkälti alusten mahdollisen kulkureitin. Lisäksi pohjoisemmassa, jossa jää on paksumpaa, on reittien ennusteet tarkempia. (Hakola 2020)

Ennustemalleja talvimerenkulkuun liittyen on tehty useita, joten tässä työssä ei ole tarkoituksena käsitellä näitä kaikkia. Alla esitettyyn taulukkoon 3 on kuitenkin koottu listaus tutkimuksista, joissa käsitellään talvimerenkulun ennustemalleja.

Kehittäjä(t)	Ennustemalli	Tutkimus
Montewka et al.	Aluksen suorituskyvyn ennustaminen	Towards probabilistic models for the prediction of a ship performance in dynamic ice
Hakola, V.	Reitin ja meriliikenteen ennustaminen	Predicting Marine Traffic in the Ice-Covered Baltic Sea
Rao et al.	Aluksen nopeuden ennustaminen	Predicting vessel speed in the Arctic without knowing ice conditions using AIS data and decision trees
Similä, M. & Lensu, M.	Aluksen nopeuden ennustaminen	Estimating the Speed of Ice-Going Ships by Integrating SAR Imagery and Ship Data from an Automatic Identification System
Vanhatalo et al.	Aluksen jäihin juuttumisen ennustaminen	Probability of a ship becoming beset in ice along the Northern Sea Route – A Bayesian analysis of real-life data
Milaković et al.	Aluksen nopeuden ennustaminen	A machine learning-based method for simulation of ship speed profile in a complex ice field
Löptien, U. & Axell, L.	Aluksen nopeuden ja jään ennustaminen	Ice and AIS: ship speed data and sea ice forecasts in the Baltic Sea

Taulu 3: Nykyisiä talvimerenkulun ennustemalleja

3. SELITETTÄVÄ TEKOÄLY

Jos joutuisit vastuuseen koneen tekemästä päätöksestä, jolla olisi merkittäviä taloudellisia, turvallisuuteen vaikuttuvia tai muita merkittäviä seurauksia, luottaisitko sokeasti tähän koneen tekemään päätökseen? (Doran et al. 2017)

Yksi tekoälyyn liittyvä keskeinen kysymys on, että miten tekoälyjärjestelmät ja niiden tekemät päätökset voidaan pitää vastuullisina. Doran et al. (2017) mukaan koneen tekemän merkittävän ja eettisen päätöksen hyväksymistä varten tulisi henkilöllä olla syvällisempää ymmärrystä järjestelmän päätöksenteon perusteista sen sijaan, että päätöksiin luotettaisiin sokeasti. Jotta täydellinen luotettavuus sekä eettiset ja moraaliset standardit voitaisiin saavuttaa, ovat tekoälypäätösten selitteet tarpeellisia. Tekoälyn tulisikin tarjota tulosten lisäksi myös ihmisen ymmärtävä selite, joka ilmaisee perusteet tuloksille, jolloin esimerkiksi analyytikot voivat selitteeseen perustuen arvioida, että perustuuko päätös rationaalisiin argumentteihin tai ovatko päätökset ristiriidassa etiikan tai lakien kanssa.

Tekoälyn selitettävyys on noussut aktiivisen tutkimuksen aiheeksi, sillä käyttäjien kokema turvallisuuden ja luottamuksen tarve on kasvanut automatisoidun päätöksenteon myötä. Automatisoidun päätöksenteon soveltamisessa, kuten autonomisessa ajamisessa, lääketieteellisissä diagnooseissa tai pankki- ja rahoitustoiminnassa kysymykset kuten miksi ja miten ovat hyvin keskeistä ymmärtää. Vaikka tekoälyn selitettävyyteen on viime vuosina kiinnitetty paljon huomiota, sen juuret ulottuvat vuosikymmenien taakse, jolloin tekoälyjärjestelmien sijaan niitä kutsuttiin asiantuntijajärjestelmiksi (*eng. expert system*). Sittemmin tekoälyn selitettävyyden määrittelyä, ymmärtämistä sekä toteutusta on tutkittu monilla eri aihealueilla, kuten asiantuntijajärjestelmissä, koneoppimisessa ja suosittelujärjestelmissä. (Confalonieri et al. 2019)

Kyky antaa selitys sille, että miksi jokin tietty päätös on tehty, on yksi nykyisten tekoälyjärjestelmien toivottava ominaisuus. Selitykset auttavat käyttäjiä ymmärtämään, ylläpitämään sekä käyttämään niitä tehokkaasti. Selitteiden tulisi myös muun muassa pystyä auttamaan käyttäjää mallien testauksessa virheellisten johtopäätösten välttämiseksi. Tekoälyn yhteydessä puhutaan myös niin sanotun mustan laatikon ongelmasta (*eng. Black box problem*), jolla tarkoitetaan tekoälyjärjestelmiä, joiden algoritmit ja toiminnot eivät ole läpinäkyviä sen käyttäjäryhmälle. Jos algoritmeja ei pystytä selittämään, ei niitä vastaan voida argumentoida, parantaa niitä eikä oppia niistä. Tähän ongelmaan selitettävä tekoäly pyrkii vastaamaan. (Confalonieri et al. 2019; Gerlings et al. 2021)

Seuraavissa alaluvuissa käydään läpi tarkemmin, miksi tekoälyn selitettävyyttä tarvitaan, erilaisia selitettävyyden tasoja sekä millaisia eri selitteitä on yleisesti käytössä. Lisäksi perehdytään itse teknologioihin, että miten tekoälyselitteitä voidaan luoda sekä mitä haasteita liittyy yleisesti ottaen selitettävään tekoälyyn.

3.1 Miksi selitettävyyttä tarvitaan?

Gerlings et al. (2021) mukaan tekoälyn selitettävyyden tarvetta voidaan perustella muun muassa seuraavilla eri hyödyillä

- Luottamuksen, läpinäkyvyyden ja ymmärryksen luonti
- Regulaatioiden ja lakien noudattamisen varmistaminen
- Sosiaalisen vastuun, oikeudenmukaisuuden ja riskien huomioiminen
- Vastuullisten, luotettavien sekä järkevien mallien luonti
- Mallien vinoumien ja väärinkäsitysten minimointi
- Mallien validointi ja tekoälyselitteiden vahvistaminen.

Gerlings et al. (2021) jatkavat, että luottamuksen luonti on selitettävän tekoälyn keskeisimpiä ajureita ja liittyy vahvasti myös tekoälyn läpinäkyvyyteen. Asiaa voidaan lähestyä kahdelta eri toisiaan täyttävällä lähestymistavalla: 1) Selitettävyys ymmärretään siten, kuinka hyvin ihminen ymmärtää selityksen tietyssä kontekstissa ja 2) ennusteen (päätöksen) selitys ihmisille (kohdeyleisölle). Tarkoittaen, että usein tekniset xAI-lähestymistavat pyrkivät saamaan tietoa esimerkiksi mallin muuttujien tärkeydestä (*eng. feature importance*) tai herkkyysanalyysilla (*eng. sensitivy analysis*) luodakseen läpinäkyvyyttä. Tällaiset lähestymistavat ja viitekehykset perustuvat pääosin läpinäkyvyyden käsitykseen ja voivat parantaa ymmärrystä ja siten lisätä luottamusta – tai päinvastaisesti vähentää luottamusta johtuen mustan laatikon ongelmasta. Lisäksi Gerlings et al. (2021) mainitsevat mallien sosioteknisen näkökulman. Yhä harvemmin ymmärretään huomioida eri sidosryhmien tarve sosioteknisille selitteille ja ihmisten sekä tietokoneiden välisen vuorovaikutuksen dilemma (*eng. HCI-dilemma*) sekä kehittäjien kehittäjille luomien selitteiden aiheuttamat riskit. Kehittäjien tai datan parissa toimivien teknisten henkilöiden tuottamat XAI-selitteet eivät siten välttämättä ratkaise luottamuskysymystä,

joten tekoälyn selitettävyyden kannalta olisi tärkeää keskittyä yhä enemmän myös ihmisten ymmärtävyyteen ja tulkittavuuteen, eikä ainoastaan läpinäkyvyyden tuottamiseen.

Uusien regulaatioiden ja GDPR-lakien avulla on mahdollista saada tekoälyn selitettävyys nostettua yksittäisten sidosryhmien lisäksi koskemaan myös koko yhteiskuntaa. Tämä edellyttää, että ammatinharjoittajat sekä teollisuus lisäävät investointeja ”läpinäkymättömien” mallien selittämiseen. GDPR-regulaatio ja ihmisten ymmärrys vaatia selitettävyyttä ovat herättäneet niin tutkijat kuin teollisuuden suuntautumaan yhä enemmän kohti selitettävää tekoälyä. Lisäksi tutkijat ovat puhuneet itse xAI:n sääntelystä tai erilaisten standardimallien käyttöönotosta, jolla voitaisiin varmistaa xAI:n vastuullinen käyttö ja pyrittäisiin välttämään ”suostuttelevien” mallien luominen selitettävien mallien sijasta. (Gerlings et al. 2021)

Kolmas Gerlings et al. (2021) mainitsema syy selitettävyyden tarpeelle on sosiaalisen vastuun, oikeudenmukaisuuden sekä riskien huomioiminen. Erityisesti terveydenhuollossa ja kliinisessä sekä oikeudellisessa työssä riskit ja vastuu ovat suuri huolenaihe, sillä ne voivat koskea ihmishenkiä eivätkä vain kustannus-hyötyanalyyseja. Riskejä voidaan vähentää, kun vastuu annetaan yksittäiselle ammattilaiselle. Tämän vuoksi on kehitetty ajatusmalleja asiantuntijapäättelylle, jotta erilaisten koneoppimismallien ja syvien neuroverkkojen taustat ymmärretään paremmin. Muun muassa läpinäkymättömien mallien syrjintä päätöksenteossa on herättänyt keskustelua mallien oikeudenmukaisuudesta sekä mallien rakenteiden syvemmästä ymmärtämisestä. Yksi selitettävän tekoälyn tärkeimmistä näkökulmista onkin varmistaa mallien oikeudenmukaisuus ja puolueettomuus auditoimalla ja luomalla todisteita niiden oikeellisuudesta.

Tekoälymallien vinoumat (*eng. bias*) ovat Gerlings et al. (2021) mukaan yksi selittävä tekijä selitettävän tekoälyn kasvuun. Esimerkiksi tiedotusvälineiden uutiset tekoälymalleja kohtaan, jotka suoriutuvat ihmisiä heikommin eri tehtävistä, kuten työnhakijoiden suodattamisesta pois palkkausprosesseissa, ovat nostaneet ennakkoluuloja tekoälyä kohden ja siksi selitettävyys on noussut tärkeäksi tekijäksi. Varsinkin neuroverkoista ja niiden harjoitusdatasta puhuttaessa vinoumilla on suuri merkitys mallin kelpoisuuteen. Jos neuroverkko luotaisiin moottoriajoneuvoja tunnistamista varten, mutta harjoitusdata sisältäisi pääosin pelkkiä autoja, tunnistaisi malli autoja todennäköisesti hyvin, mutta muita ajoneuvoja huonosti. Tällöin puhuttaisiin vinoutuneesta datasta, joka johtaa myös vinoutuneeseen malliin. Selitettävä tekoäly parantaa mallien läpinäkyvyyttä ja täten mahdollistaa myös edellä esitetyn kaltaisten vinoumien huomaamisen.

Adadi & Berrada (2018) perustelevat selitettävän tekoälyn tarvetta samoin perustein kuin edellä esitetty: perustelemisen, kontrollin, kehittymisen sekä uusien asioiden löytämisen vuoksi. Kun puhutaan päätösten selittämisestä, tarkoitetaan sillä yleensä syiden tai perustelujen tarvetta saadulle tulokselle, eikä päätöksentekoprosessin sisäisen toiminnan tai päättelyn logiikan kuvausta. XAI-järjestelmien käyttö tarjoaa tarvittavat keinot näiden tulosten perustelemiseksi, varsinkin kun tehdään odotusten vastaisia päätöksiä. Järjestelmät myös varmistavat, että on olemassa oikeudenmukaisesti sekä eettisesti auditoitavia ja todistettavissa olevia tapoja puolustaa algoritmien tekemiä päätöksiä, joka parantaa luottamusta. (Adadi & Berrada, 2018)

Päätösten perustelemisen lisäksi tekoälyn selitettävyys voi myös auttaa estämään asioita menemästä pieleen ja siten tehostamaan kontrollia. Järjestelmän käyttäytymisen ymmärtäminen mahdollistaa tuntemattomien haavoittuvuuksien ja puutteiden paremman näkyvyyden sekä auttaa tunnistamaan ja korjaamaan virheet nopeasti. Selitettävyys on lisäksi keino kehittää ja parantaa malleja jatkuvasti, koska mallia, jota pystytään selittämään ja ymmärtämään, on helpompi myös parantaa. Kun käyttäjät tietävät miksi järjestelmä tuottaa saatuja tuloksia, on siitä mahdollista kehittää älykkäämmäksi. Voidaan siis sanoa, että selitettävä tekoäly voi olla perustana tekoälyn jatkuvalle iteraatiolle ja kehittymiselle. Selitteiden avulla on myös mahdollista oppia uusia asioita, kerätä tietoa sekä siten luoda uudenlaista ymmärrystä. Esimerkiksi, jos peleissä ihmistä älykkäämpi tekoäly pystyisi selittämään oppimansa pelistrategiansa, voisi se mahdollistaa uudenlaisen tietämyksen syntymisen. (Adadi & Berrada, 2018)

3.2 Tekoälyn selitettävyyden tasot

Selitettävä tekoäly (XAI) on laaja käsite ja eri sidosryhmät voivat ymmärtää tekoälyn selitettävyyden tarpeen eri tavoin, minkä vuoksi tässä työssä tarkastellaan selitettävää tekoälyä eri tasoista. Kirjallisuudessa käytetään erilaisia selitettävyyden tasoja sidosryhmille, eikä yhtä oikeaa tapaa jakaa näitä ole, mutta tässä työssä tutkitaan asiaa Preece et al. (2018) ja Hong et al. (2020) esittämien selitettävän tekoälyn sidosryhmien pohjalta, joita ovat kehittäjät, teoreetikot ja asiantuntijat, eetikot sekä käyttäjät.

3.2.1 Kehittäjät

Kehittäjillä viitataan niihin henkilöihin, jotka ovat tekemisissä tekoälyratkaisujen rakentamisen kanssa. Monet tähän tasoon liittyvistä henkilöistä työskentelevät esimerkiksi teollisuudessa tai julkisella sektorilla ja luovat tekoälyratkaisuja monista eri syistä, ku-

ten auttaakseen heitä omassa työssänsä. Kehittäjäyhteisössä käytetään selitettävyyden lisäksi myös tulkittavuuden (*eng. interpretability*) termiä ja muun muassa kehittäjien yksi motiivi tekoälyn selitettävyyden ja tulkittavuuden tarpeelle on laadunvarmistus eli tekoälyratkaisujen testauksen, virheenkorjauksen ja arvioinnin tukeminen ja näiden ratkaisujen kestävyyden parantaminen. Bhatt et al. (2020) lisäävät, että tekoälyratkaisujen kehittäjät mukaan lukien datatieteilijät ja tutkijat kehittävät koneoppimismalleja ja käyttävät selitteitä mallien toiminnan ymmärtämiseksi. Kehittäjät voivat käyttää selitteiden luomiseen kehitettyjä avoimen lähdekoodin kirjastoja, joista laajasti käytettyjä ovat muun muassa LIME, SHAP, Deep Taylor Decomposition ja erilaiset vaikutusfunktiot (*eng. Influence functions*) (Preece et al. 2018). Luvussa 3.4. perehdytään syvällisemmin teknologioihin selitteiden takana.

Hong et al. (2020) mukaan tutkimuksia, joiden tarkoituksena on empiirisesti ymmärtää, kuinka kehittäjät käytännössä katsoen kokevat selitettävyyteen ja tulkittavuuteen liittyviä tehtäviä sekä millaisia kehittäjien käytännöt, tarpeet ja haasteet ovat asian suhteen, on vielä ollut suhteellisen vähän. Tulkittavuus liitetään usein siihen, kuinka hyvin malli kommunikoi päätöksensä käyttäjille, mutta paljon vähemmän tiedetään, kuinka tulkittavuus ilmenee käytännössä työpaikoilla, joissa kehittäjien on kommunikoitava ja koordinoitava työtään mallien sekä päätöksentekotyökalujen ympärillä. Tämän ongelman ratkaiseminen on keskeistä, jotta ammatinharjoittajien kohtaamia todellisia ongelmia tulkittavuuskysymyksiin liittyen pystyttäisiin tutkimaan paremmin. (Hong et al. 2020)

Hong et al. (2020) tutkimuksessa selvitettiin tulkittavuutta muun muassa koneoppimismallien rakentajien työssä. Tutkimuksessa rakentajilla viitataan henkilöihin, jotka vastuussa mallien suunnittelusta, kehittämisestä ja testauksesta sekä niiden integroimisesta organisaation datainfrastruktuuriin ja yleisimpiä työnimikkeitä kehittäjien alla tutkimuksessa ovat datatieteilijät sekä datainsinöörit. Tutkimuksen mukaan tulkittavuus on merkittävässä roolissa jo heti mallin suunnitteluvaiheesta lähtien ennen kuin itse mallia on kehitetty ja mallin ominaisuuksien suunnittelu yhdessä päättäjäryhmän kanssa helpottaa heidän luottamuksensa saamista varmennus- ja validointivaiheissa. Hong et al. (2020) jatkavat, että kehittäjät lähestyvät ongelmaa tyypillisesti kolmen eri ”tulkittavuuden linssin” läpi, joita ovat tapaukset, muuttujat ja mallit. Tapauslinssillä tarkoitetaan yksittäisten tapausten tai pienten tapausjoukkojen tutkimista. Yksi yleinen validointistrategia on luoda testitapauksia, jossa määritetään ja tutkitaan tapausjoukkoja, joissa mallien tulisi toimia odotetulla tavalla. Tällaiset tapaukset luodaan yleensä yhdessä henkilöiden kanssa, joilla on syvällisempää tietoa toimialasta sekä liiketoiminnan tarpeista. Olennainen asia käytettäessä testitapauksia validoinnissa on ymmärtää miten ja miksi

malli tuottaa tietyn tulosteen testitapauksessa - varsinkin silloin, jos malli ei toimi odotetusti. Tähän aikaisemmin mainitut työkalut, kuten SHAP ja LIME voivat olla avuksi.

Muuttujiin keskittyminen auttaa havaitsemaan, mitkä muuttujat ohjaavat eniten mallin päätöksiä. Muuttujien tärkeyttä analysoitaessa on kuitenkin huomioitava, että vähiten tärkeiden muuttujien tarkastaminen voi olla yhtä tärkeää, ellei jopa tärkeämpää, kuin tärkeimpien muuttujien tarkastelu. Kun tietoa muuttujien tärkeydestä jaetaan muiden sidosryhmien kanssa, auttaa se kollektiivisesti ymmärtämään mallin käyttäytymistä. Kehittäjien useiden eri mallien vertailu on yksi keskeisimmistä vaiheista, sillä mallin tulkittavuus ei synny yhdestä mallista, vaan enemmänkin usean mallin sarjasta, jota kehitetään ja parannetaan asteittain.

3.2.2 Teoreetikot ja toimialaosaajat

Teoreettisen ajattelun tasolle kuuluvat henkilöt, jotka ovat kiinnostuneita tekoälyn teorian ymmärtämisestä ja edistämisestä etenkin syvien neuroverkkomallien ympärillä. Tähän tasoon liittyvät henkilöt ovat pääosin akateemisista tai teollisuuden tutkimusyksiköistä ja teoreetikkotaso eroaakin kehittäjätasosta siten, että heidän tärkein motivaationsa on edistää tekoälyä entistä paremmaksi käytännön sovellusten tekemisen sijaan. Tavoitteena on ymmärtää tekoälyn perusominaisuuksia ja siten teoreettisella tasolla puhutaankin enemmän tulkitsevuudesta kuin selitettävyydestä. Teoreettinen taso on vahvasti sidoksissa kehittäjätason kanssa. Esimerkiksi tutkija, joka tutkii teoreettista puolta tekoälymallin tekniikasta (teoreetikko) soveltaa myös tekniikkaa mallin luontiin (kehittäjä). Teoreetikot ovat siten kehittäjien tavoin tekoälyratkaisujen luojia. (Preece et al. 2018)

Hong et al. (2020) jatkavat, että kehittäjät usein hyödyntävät toimialaosaajia varmistaakseen, että heidän tekemänsä valinnat eivät vaikuta negatiivisesti mallin tulkittavuuteen ja mahdollistavat oikeanlaisten ominaisuusjoukkojen suunnittelun. Toimialaosaajat ovat henkilöitä, joilla on tietoa toimialasta varmistaakseen, että mallit täyttävät halutut tavoitteet sekä käyttäytyvät odotetulla tavalla, mutta eivät omista ammatillista osaamista tekoälystä. Esimerkiksi liiketoiminta-analyytikot on hyvä ottaa mukaan suunnitteluun jo varhaisessa vaiheessa, kun määritellään luotonarviointimalleja, jotta malli vastaa tulkittavuuden ja selitettävyyden tarpeisiin heti alusta alkaen.

3.2.3 Eetikot

Eettisellä tasolla henkilöt ovat kiinnostuneita tekoälyratkaisujen oikeudenmukaisuudesta, vastuullisuudesta ja läpinäkyvyydestä. Vaikka eettisellä tasolla on usein tietojenkäsittelytieteilijöitä sekä insinöörejä, se on laajalti monialainen taso sisältäen muun muassa yhteiskuntatieteilijöitä, lakimiehiä, toimittajia, taloustieteilijöitä sekä poliitikkoja. Selitettävyyden ja tulkinnallisuuden lisäksi myös ymmärrettävyys sekä luotettavuus ovat tärkeitä näkökulmia eetikoille. Eettiselle tasolle kuuluvat henkilöt voivat kuulua myös kehittäjätasolle sekä teoreettiselle tasolle, mutta motiivit selitysten tarpeille ovat erilaiset. Tällä tasolla selitteiden on ulotuttava teknisen ohjelmiston laadun ulkopuolelle, jotta voidaan taata muun muassa oikeudenmukaisuus, puolueeton käyttäytyminen sekä avoimuus vastuullisuuden ja tarkastettavuuden mahdollistamiseksi mukaan lukien lainsäädännön noudattaminen, kuten Euroopan Unionin GDPR-tietosuoja-asetus. (Preece et al. 2018)

Launis (2020) mukaan Euroopan Union yleinen tietosuoja-asetus (GDPR) vaatii, että yritys, joka hyödyntää henkilötietoja tekoälyjärjestelmän automaattisessa käsittelyssä, on pystyttävä selittämään miten kyseinen järjestelmä tekee päätöksiä. Talvimerenkulun tekoälyratkaisuissa henkilökohtaisia henkilötietoja ei ole tarpeen käyttää, joten GDPR-asetuksella ei suoraan ole talvimerenkuluun vaikutusta. Euroopan Unioni on kuitenkin julkaissut hiljattain uuden ehdotuksen tekoälyä sääntelevästä asetuksesta (Viljanen 2021), joten varsinaisiin tekoälyratkaisuihin on tulevaisuudessa kiinnitettävä entistä tarkempaa huomiota. Viljanen (2021) kertoo, että ehdotus koskee vain korkean riskin tekoälyjärjestelmiä. Ehdotuksessa tekoälyjärjestelmä määritellään ratkaisuna, jossa datalähteet muuttuvat tulosteiksi, mutta korkean riskin tekoälyjärjestelmät ovat esitetty neljässä eri kategoriassa. Ensimmäiseen kategoriaan kuuluvat muun muassa ihmisten käyttäytymiseen vaikuttavat järjestelmät, poliisien kasvojentunnistus- sekä sosiaaliset luokitusjärjestelmät. Toiseen kategoriaan tuoteturvallisuussääntelyyn kuuluvat tuotteet ja kolmanteen kategoriaan muun muassa ajoneuvot, lentokoneet ja laivat, joten myös talvimerenkulun tekoälyratkaisut kuuluisivat korkean riskin AI-järjestelmiin. Neljänteen kategoriaan kuuluvat esimerkiksi biometriseen tunnistamiseen, lainvalvontaan ja koulutukseen liittyvät järjestelmät. (Viljanen, 2021)

Ehdotuksessa on esitetty kuusi eri sääntelytapaa tekoälyjärjestelmille, joita ovat kieltosääntely, johtamisjärjestelmäsääntely, sitova teknologiasääntely, etukäteinen valvonta, jälkikäteinen valvonta sekä sanktiot. Kieltosääntelyllä voidaan kieltää kokonaan tietynlaiset järjestelmät, kuten manipulatiiviset järjestelmät, joilla pyritään vaikuttamaan yksi-

löiden toimintaan ja johtamisjärjestelmäsääntely puolestaan pyrkii vaikuttamaan suoraan liikkeenjohdollisiin keinoihin, kuten riskienhallintaan ja laadunvarmistukseen sekä datan hallintaan. Tällä tavoin yritykset ovat itse paremmin tietoisia tekemisistään muun muassa dokumentoimalla tehtyä työtä. Sitovan teknologiasääntelyn avulla pyritään puolestaan sääntelemään, millaisia teknisiä ratkaisuja tekoälyjärjestelmissä on käytettävä. Etukäteisvalvonnassa nimensä mukaisesti valvotaan AI-järjestelmiä etukäteen ja järjestelmä voi esimerkiksi tarvita CE-merkinnän, jotta se voidaan tuoda markkinoille tai ottaa käyttöön sisäisesti. Jälkikäteisvalvonnan avulla toisaalta markkinavalvontaviranomaiset voivat tarkkailla järjestelmiä ja jos ne eivät täytä määräyksiä, voidaan vaatia vaatimusten täyttämistä tai vetämistä pois markkinoilta. Viimeisenä sääntelytapana ovat sanktiot, jotka voivat suurimmillaan olla jopa kuusi prosenttia liikevaihdosta, jos esitettyjä sääntöjä rikotaan. (Viljanen, 2021) Voidaan siis todeta, että tekoälyn selitettävyyden eettisen tason merkitsevyys on regulaatioiden ja säännösten myötä nousemassa vuosi vuodelta entistä tärkeämmäksi osaksi tekoälyratkaisujen kehittämistä.

3.2.4 Käyttäjät

Käyttäjätasolle kuuluvat nimensä mukaisesti tekoälyratkaisuja käyttävät henkilöt, jotka tarvitsevat selkeitä ja ymmärrettäviä selitteitä tekoälymallista, jolloin selitettävyystarpeet ovat luonnollisesti erilaisia kuin esimerkiksi kehittäjillä. Kolme edellä esitettyä tasoa muodostavat valtaosan henkilöistä, jotka osallistuvat tekoälyn selitettävyyttä käsittelevän kirjallisuuden ja materiaalin luomiseen. Käyttäjät puolestaan eivät, sillä he tarvitsevat selitteitä, joiden avulla he kykenevät päättämään, miten toimia tekoälyn tulosten perusteella ja auttaa perustelemaan nämä toimet. Käyttäjätasolle kuuluvat loppukäyttäjät, mutta lisäksi kaikki tekoälyjärjestelmän vaikuttamiin prosesseihin kuuluvat henkilöt (Preece et al. 2018). Hong et al. (2020) lisäävät, että käyttäjätasolle kuuluu laajasti eri ammattilaisia, kuten lääkäreitä, pankkien edustajia tai biologeja, jotka pyrkivät tekemään korkean tason päätöksiä tekoälymallien tukemana.

Jin et al. (2021) kehittämä EUCA-viitekehys (*End-User-Centered Explainable AI Framework*) pyrkii vastaamaan kokonaisvaltaisesti selitettävän tekoälyn haasteisiin loppukäyttäjän näkökulmasta. Viitekehys koostuu neljästä eri selitealueesta, joita ovat ominaisuuksiin, esimerkkeihin ja sääntöihin perustuvat selitteet sekä yleistiedot mallista. Ominaisuuksiin pohjautuvissa selitteissä loppukäyttäjälle pyritään antamaan tietoa ominaisuuksien tärkeydestä ja vuorovaikutuksesta toisien ominaisuuksien kanssa selkeästi visuaalisesti esitettynä. Ihmisillä on tapana käyttää esimerkkejä oppimiseen ja asioiden selittämiseen, joten esimerkkeihin pohjautuvat selitteet ovat loppukäyttäjälle helposti

tulkittavia. Tällaisia esimerkkejä ovat muun muassa samanlaiset, tyypilliset sekä vaihtoehtoiset esimerkit. Sääntöihin perustuvat selitteet ovat selitteitä, joissa mallin päätökset voidaan joko kokonaan tai osittain kuvata loogisilla ehtolauseilla. Loppukäyttäjälle on myös tärkeää antaa yleistietoa mallista, kuten syöte, tuloste, suorituskyky sekä käytetty data. Seuraavassa luvussa 3.3 käydään tarkemmin läpi lisää erilaisia tekoälyselitteitä.

Loppujen lopuksi keskeisintä loppukäyttäjän näkökulmasta on, että selitettävyys ja tulkittavuus on sillä tasolla, että se antaa loppukäyttäjälle mahdollisuuden entistä parempaan päätöksentekoon. Päätöksentekijät eivät vain voi luottaa ennusteisiin, saati käyttää niitä, jos ennusteita ei olla perusteltu riittävän syvällisesti. Samoin selitteiden on vastattava loppukäyttäjien tarpeita sekä toimialan rajoituksia. (Hong et al. 2020)

Koska jokaisella edellä mainitulla sidosryhmällä on oma yksilöllinen ymmärrys järjestelmästä, yksilöllinen kyky ymmärtää ja yksilöllinen joukko toiveita, jotka täytyy jossain määrin täyttää, selitettävää tietoa saavien sidosryhmien ominaisuudet vaikuttavat selitettävyyden ja sen ymmärtämisen väliseen suhteeseen. Tällaisia ominaisuuksia on muun muassa osaaminen, uskomukset, oppimiskyky ja toiveet järjestelmiä kohtaan. Esimerkiksi teknisiä yksityiskohtia sisältävät selitteet voivat lisätä asiantuntevan kehittäjän ymmärrystä, kun taas tekniset yksityiskohdat voivat haitata muiden ei teknisten sidosryhmien ymmärrystä. (Langer et al. 2021)

Langer et al. (2021) jatkavat, että kullekin sidosryhmälle keskeiset toiveet voivat vaikuttaa selitettävän tiedon ja ymmärryksen väliseen suhteeseen. Erityisesti, kun sidosryhmät käyttävät selitettävyyttä edistääkseen ymmärrystä tekoälyjärjestelmistä, heidän motivaationsa sekä aiemmat uskomukset voivat vaikuttaa siihen, miten he tulkitsevat saatua selitettä. Jos sidosryhmän ensisijaisena toiveena on esimerkiksi varmistaa, että tekoälyjärjestelmä tarjoaa oikeudenmukaisia tuloksia, saattavat he keskittyä erilaisten vinoumien tarkasteluun, jotka voisivat johtaa epäoikeudenmukaisiin tuloksiin. Toisaalta, jos sidosryhmän ensisijaisena toiveena on parantaa järjestelmän ennustetarkkuutta, voivat sidosryhmään kuuluvat henkilöt kiinnittää erityistä huomiota tietoihin, jotka tarjoavat näkemyksiä järjestelmän suorituskyvyn parantamisesta. Voidaan todeta, että toiveesta riippuen sama selite voi johtaa eriasteisiin ja erilaiseen ymmärrykseen. Tästä johtuen on tärkeää antaa oikeanlaista selitystä haluttuun tarkoitukseen.

Edellä esitettyjä sidosryhmiä voidaan nähdä myös talvimerenkulkuun liittyvän tekoälyn kehittämisessä ja varsinkin talvimerenkulun luoman haastavuuden vuoksi eri sidosryhmien huomioiminen on tärkeää. Talvimerenkulussa jääolosuhde sekä monet muut muuttujat, kuten tuulen vaikutukset tekevät kehittäjille mallien kehittämisen ja esimerkiksi SHAP-selitteiden tulkitsemisen haastavaksi, jolloin teoreetikot ja toimialaosaajat

voivat tukea selitteiden tulkitsemisessa ja siten mallin kehittämisessä. Tekoälyratkaisujen lisääntyessä tulevat myös näiden säännöstely todennäköisesti lisääntymään ja se tulee koskemaan yleisesti myös laivaliikennettä. Tämän vuoksi eetikoiden huomioiminen selittäessä on tärkeää, jotta erilaisiin asetusten ja regulaatioiden vaatimiin läpinäkyvyys- ja vastuullisuuskysymyksiin talvimerenkulussa voidaan vastata. Loppukäyttäjät, kuten aluksen kapteeni ja muu henkilökunta tarvitsevat puolestaan helposti tulkittavia selitteitä, joista käy selkeästi esimerkiksi mitä tulee mahdollisesti tapahtumaan ja mihin se perustuu. Jos kehitysprosessissa otetaan selitettävyys huomioon jokaisessa sidosryhmässä heti prosessin alusta lähtien, mahdollisuudet läpinäkyvämmän ja laadukkaamman tekoälyratkaisun kehittämiselle parantuvat.

3.3 Erilaiset tekoälyselitteet

Tässä työssä tekoälyselitteitä tarkastellaan Preece et al. (2018) mainitseman kolmen eri selitetyypin pohjalta. Näitä ovat läpinäkyvyyteen perustuvat selitteet, post-hoc -selitteet sekä kerrostetut selitteet, jotka käydään läpi seuraavissa alaluvuissa.

3.3.1 Läpinäkyvyyteen perustuvat selitteet

Läpinäkyvä malli on niin sanotun mustan laatikon vastakohta ja se merkitsee jonkinlaista ymmärrystä mallin toimintamekanismista. Lipton et al. (2017) jakavat läpinäkyvyyden koko mallin tasolle (simuloitavuus), yksittäisten komponenttien tasolle (pilkottavuus) sekä opetusalgoritmien tasolle (algoritminen läpinäkyvyys), jolloin kokonaisvaltainen läpinäkyvyys vaatisi läpinäkyvyyttä jokaisella tasolla. Tarkimmassa tapauksessa mallia voidaan kutsua läpinäkyväksi, jos sitä tutkiva henkilö pystyy tutkimaan koko mallia kerralla. Komponenttitasolla jokainen syöte, parametri ja laskutoimitus sisältää intuitiivisen selitteen ja algoritmitasolla ymmärretään opetusalgoritmeja syvällisesti, kuten esimerkiksi lineaaristen mallin virhepinnan muotoa, joita tavallinen käyttäjä ei osaa tulkita. Kokonaisvaltaisen läpinäkyvyyden saavuttaminen on kuitenkin lähes mahdotonta muille kuin pienille ja yksinkertaisille malleille, kuten yksinkertaisille päätöspuille. (Lipton et al. 2017)

Preece et al. (2018) mukaan esimerkiksi näkyvyyskartan (*eng. saliency map*) kaltaiset karttaselitteet, jotka kertovat tulosteen kannalta merkityksellisimmistä ominaisuuksista, voivat tuoda välitöntä arvoa kehittäjille sekä teoreetikoille. Tällaiset tekniikat, jotka visualisoivat esimerkiksi aktivointeja neuroverkon syöte- tai piilokerroksissa ovat kuitenkin läpinäkyvyydeltään rajoitetumpia. Vaikka visualisoinnin elementti näissä lähestymista-

voissa viittaavat Post-hoc-tekniikkaan, on kyseessä kuitenkin Preece et al. mukaan läpinäkyvyyteen perustuvat selitteet. Läpinäkyvyyteen perustuvat attribuuttivisualisoinnit voivat olla lisäksi käyttäjille ja eetikoille vaikeasti tulkittavia, jos selitteet eivät näissä tuo esiin syötteen merkityksellisiä piirteitä. Siksi läpinäkyvyyteen perustuvat selitteet voivat tehdä näistä sidosryhmistä vähemmän taipuvaisia luottamaan järjestelmään. Vaikka selite vaikuttaisi vakuuttavalta sen korostaessa merkityksellisiä ja uskottavia piirteitä, on olemassa kuitenkin niin sanotun vahvistusharhan vaara, ellei muita vaihtoehtoisia tapauksia huomioida. Yksityiskohtaisten läpinäkyvyyteen perustuvien selitteiden tarjoaminen voi myös olla vastaanottajalle liiallista, jolloin suurempi määrä tietoa voi olla huonompi käyttäjän suorituskyvyn kannalta.

3.3.2 Post hoc -selitteet

Post hoc on muista erottuva lähestymistapa hankkia tietoa käytetystä mallista ja sen oppimisesta. Se käyttää syötteenä opetettua ja/tai testattua tekoälymallia ja luo sen pohjalta hyödyllisiä approksimaatioita mallin sisäisestä toiminta- ja päätöksentekologiikasta tuottamalla siitä ymmärrettäviä esityksiä. Monet post hoc -menetelmät pyrkivät tuomaan esille ominaisuuksien arvojen ja tulosteen välisiä yhteyksiä. Tämä auttaa käyttäjiä tunnistamaan ja mittaamaan tärkeimmät ominaisuudet sekä tunnistamaan mahdolliset vinoumat datassa ja käytetyssä mallissa. (Moradi & Samwald, 2021)

Vaikka post hoc -selitteet eivät usein kerro tarkasti, miten malli toimii, ne voivat silti tarjota hyödyllistä neuvoa koneoppimismallien kanssa työskenteleville sekä niiden loppukäyttäjille. Yleisiä post-hoc -selitteitä ovat muun muassa luonnollisen kielen selitteet, mallien visualisoinnit, paikalliset selitteet tai esimerkkeihin perustuvat selitteet. Tekstiselitteet toimivat usein hyvin, sillä ihmiset usein perustelevat päätöksiä suullisesti. Siten voidaan kehittää esimerkiksi yksi malli luomaan ennusteita ja toinen erillinen malli luomaan selite kyseiselle mallille. Toinen yleinen lähestymistapa post hoc -selitteiden tuottamiseen on visualisointi, jolla voidaan pyrkiä kvalitatiivisesti määrittämään mitä malli on oppinut. Esimerkiksi konenäössä visualisointi auttaa selvittämään, mitä informaatiota eri neuroverkon kerroksissa säilytetään, kuten Mahendran & Vedaldi (2015) visualisoivat muun muassa eri konvoluutioverkon kerroksia, jolloin jokaisesta kerroksesta näkee visuaalisesti mitä informaatiota kukin kerros sisältää. (Lipton, 2017; Preece et al., 2018)

Neuroverkkojen oppimista voi olla haastavaa kuvailla ytimekkäästi, joten on myös mahdollista kuvailla niitä paikallisesti. Paikalliset selitteet voivat olla kuitenkin harhaanjohtavia, sillä ne keskittyvät vain yhteen osaan mallista ja esimerkiksi eri näkyvyyskarttojen

kohdalla yhden pikselin siirtäminen voi luoda hyvin erilaisen kartan. Tämä eroaa lineaarisista malleista, jotka mallintavat syötteen ja tulosteen välisiä globaaleja suhteita. Esimerkein selittäminen on perinteinen asiantuntijoiden käyttämä tapa ja hyvä etenkin eetikko- ja käyttäjätasoille. Post hoc -tekniikat luovat selitteitä, jotka vaikuttavat samoilta kuin läpinäkyvyyteen perustuvien tekniikoiden luomat selitteet, ja jos niitä tarjotaan eetikoille tai käyttäjille, on tärkeää ilmoittaa selkeästi, että kyseessä on post hoc -selite. (Lipton, 2017; Preece et al., 2018) Rudin (2019) näkee kuitenkin post hoc -selitteiden käytössä ongelmia, sillä ne eivät ole hänen mukaansa luotettavia ja voivat olla lisäksi harhaanjohtavia.

3.3.3 Kerrostetut selitteet

Preece et al. (2017) ehdottavat kolmanneksi selitetyypiksi kerrostettuja selitteitä. Tässä selitteet ovat jaettu kolmeen eri kerrokseen, jolloin näistä pyritään yhdessä luomaan useamman sidosryhmän tarpeet tyydyttävä ”yhdistelmäseliteobjekti” hyödyntäen edellä esitettyjä selitetyyppejä. Kerroksiin kuuluvat jäljitettävyys- (*eng. traceability*), perustelu- (*eng. justification*) sekä vakuutuskerrokset (*eng. assurance*). Ensimmäinen jäljitettävyyden kerros on tarkoitettu kehittäjille sekä teoreetikoille ja se pohjautuu läpinäkyvyyteen perustuville selitteille, jotta saadaan ymmärrys mallin sisäisestä toiminnasta. Toisen tason eli perustelun pääasiallisia sidosryhmiä ovat kehittäjät ja käyttäjät. Tässä tasossa post hoc -esitykset ovat linkitetty ensimmäiseen tasoon ja tarjoaa semanttisia yhteyksiä syötteen ja tulosteen välillä osoittaakseen, että järjestelmä toimii oikein.

Kolmas taso pyrkii vastaamaan käyttäjien sekä eetikoiden tarpeisiin ja on linkitetty edeltävään tasoon. Tässä tasossa viitataan erilaisiin linjauksiin ja ontologisiin elementteihin, jotka ovat tarpeellisia, jotta vastaanottaja saa luottamusta siihen, että järjestelmä toimii oikein. Preece et al. (2017) havainnollistavat tätä esimerkin avulla: Ensimmäisellä tasolla luodaan näkyvyyskartta neuroverkon syötekerroksen piirteistä halutun kohteen luokittelua varten. Toisessa tasossa ilmoitetaan semanttisesti merkittävät piirteet eli mitä merkittäviä piirteitä malli tunnistaa esimerkiksi ihmisestä, että se kykenee tunnistamaan ihmisen. Viimeisellä tasolla voidaan antaa vaihtoehtoisia esimerkkejä, jotka näyttävät, ettei malli sekoita ihmisiä esimerkiksi eläimiin.

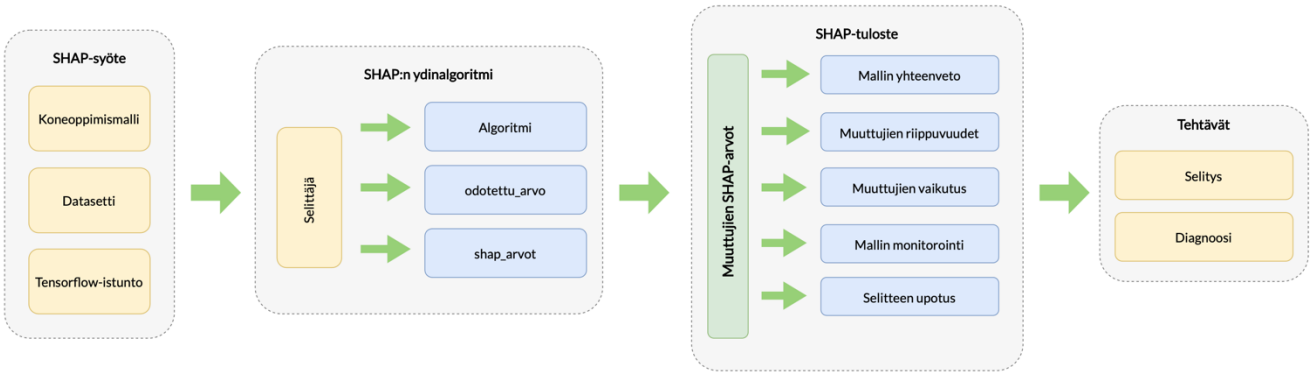
Kerrostetut selitteet toimivat talvimerenkulun yhteydessä hyvin, sillä niiden avulla kehitysprosessista saa kokonaisvaltaisen käsityksen. Muun muassa SHAP- ja LIME-menetelmillä saadaan ymmärrys mallin sisäisestä toiminnasta, kuten muuttujien tärkeydestä, mutta huomioidaan lisäksi myös aluksen käyttäjien tarpeet tarjoamalla esimerkiksi visuaalisia sekä tekstipohjaisia selitteitä.

3.4 Teknologiat selitteiden takana

Viime vuosina on esitelty monia erilaisia lähestymistapoja läpinäkyvien ja selitettävien mallien rakentamiseen mustan laatikon mallien välttämiseksi. Muun muassa PD-kuvaajat (*Partial depence plot*), ALE-kuvaajat (*Accumulated local effects*), ICE (*Individual Conditional Expectation*), SHAP-arvot (*SHapely Additive ExPlanations*) sekä LIME (*Local interpretable model-agnostic explanations*) ovat suosittuja menetelmiä mallien selittämiseen. Jokaisella edellä mainituista menetelmistä on kullakin omat lähestymistapansa, mutta ne voidaan luokitella luonnostaan läpinäkyviin, jotka ovat luonteeltaan yksinkertaisempia, mutta vähemmän tarkkoja kuin edistyneemmät mallit, kuten lineaarinen ja logistinen regressio sekä päätöspuut tai post hoc -luonteisiin agnostisiin XAI-viitekehyksiin, jotka ovat suunniteltu sopivaan mihin tahansa mallityyppiin. Tällaiset pohjautuvat tekniikkoihin, jotka muun muassa yksinkertaistavat mallia, arvioivat muuttujien tärkeyttä (SHAP), visualisoivat mallia (ICE) tai luovat paikallisen surrogaattimallin mallin tuotosta (LIME). Yhteistä näille kuitenkin on, että ne yleensä tuottavat jonkinlaisen visuaalisen näkymän ymmärtämisen helpottamiseksi. (Gerlings et al., 2021) SHAP (Lundberg & Lee, 2017) ja LIME (Ribeiro et al. 2016) ovat kaksi suosittua menetelmää selitettävään tekoälyyn liittyen, joten tässä työssä tutkitaan tarkemmin näitä XAI-menetelmiä.

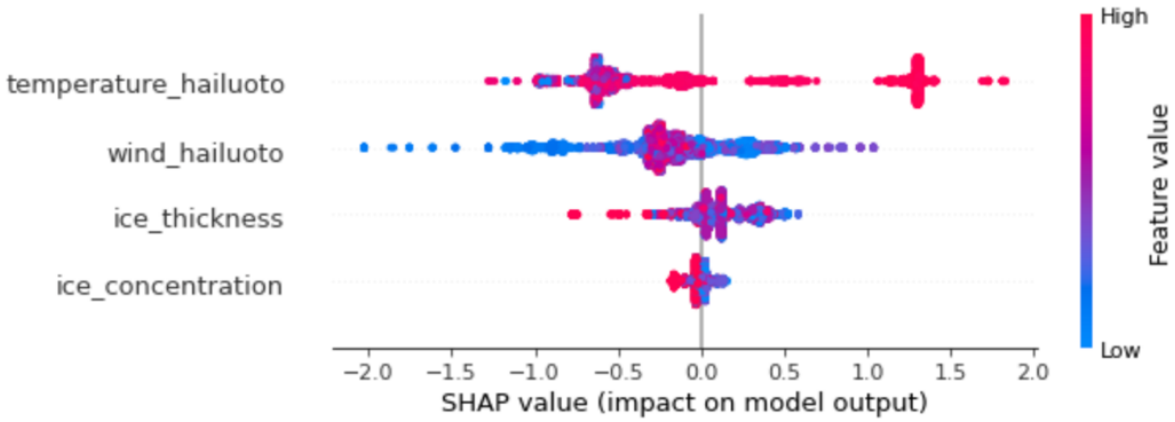
3.4.1 SHAP

SHAP on peliteoreettinen lähestymistapa minkä tahansa koneoppimismallin tulosten selittämiseen. Se yhdistää optimaalisen pisteiden allokoinnin paikallisiin selityksiin käyttäen klassisia peliteoriaan pohjautuvia Shapley-arvoja. Itse SHAP-selitysten luomiseen käytetään Python-kirjastoa, jolla pystytään hyvin yksinkertaisesti luomaan selitteitä koneoppimismalleille. Jotta SHAP-menetelmää voidaan tulkita, tulisi ymmärtää mitä Shapley-arvoihin pohjautuvat SHAP-arvot tarkoittavat. Peliteoriaan viitaten voidaan ajatella, että peli on se, joka luo uudelleen mallin tuloksen ja pelaajat ovat mallin ominaisuuksia. Shapley-arvojen voidaan siten ajatella mittaavan jokaisen pelaajan eli ominaisuuden panosta peliin ja SHAP-arvojen puolestaan määrittävän kunkin ominaisuuden vaikutuksen mallin tekemiin ennusteisiin. (Lundberg, 2018; Lundberg & Lee, 2017) SHAP menetelmässä on useita erilaisia selittäjiä (*eng. explainer*) eri malleille, kuten päätöspuille ja lineaarisille malleille. Päätöspuuselittäjä on nopea ja tarkka menetelmä SHAP-arvojen arvioimiseen erilaisille puumalleille. Gradienttiselittäjä puolestaan voi selittää mallia käyttäen oletettuja gradientteja. (Zhang et al., 2020)



Kuva 3: SHAP-viitekehys (Mukaillen Zhang et al. 2020)

Kuvassa 3 on esitetty Zhang et al. (2020) SHAP-menetelmään perustuva selitettävän tekoälyn viitekehys, joka koostuu neljästä eri vaiheesta. Ensimmäinen vaihe koostuu itse koneoppimismallista ja sen käyttämästä datasta. Tämän jälkeen malli syötetään valitulle selittäjäalgoritmille, kuten päätöspuuselittäjälle ja laskee ominaisuuksien vaikutuksen malliin. Odotetuilla arvoilla tarkoitetaan ominaisuuksien keskiarvoa. SHAP-mallin tuotos sisältää esimerkiksi mallin yhteenvedon, ominaisuuksien riippuvuudet ja vuorovaikutukset sekä mallin monitoroinnin, joista saadaan tehtäviksi selite ja diagnoosi.



Kuva 4: Esimerkki SHAP-yhteenvetokuvaajasta.

Yllä olevassa kuvassa 4 on esitetty yksinkertainen SHAP-kuvaaja neljällä eri muuttujalla: lämpötila, tuuli, jään paksuus sekä jään konsentraatio. Ennustettava arvo on laivan nopeus. Yksinkertaisuudesta huolimatta kuvaaja pitää sisällään paljon tietoa muuttujien tärkeydestä. Muuttujan sijainti kertoo sen tärkeydestä ennusteessa eli mitä ylempänä se on, sitä suurempi vaikutus sillä on malliin. Tässä tapauksessa siis lämpötilalla olisi suurin vaikutus mallin ennusteeseen ja jään konsentraatiolla pienin. Vaakasuoralla rivillä olevat pisteet kertovat negatiivisesta ja positiivisesta vaikutuksesta ennusteeseen ja värit puolestaan alkuperäisen arvon suuruudesta. Korkeammalla lämpötilalla olisi kuvaajan mukaan siis positiivinen vaikutus laivan nopeuteen. Jään konsentraatiolla puolestaan on matala vaikutus yleisesti mallin toimintaan, mutta mitä korkeampi jään konsentraatio on, sitä negatiivisempi vaikutus sillä on nopeuteen. Molnar (2022) mukaan

SHAP-menetelmä voi vaatia paljon laskentatehoa ollen siten hidas sekä myös mahdollistaa tarkoituksella harhaanjohtavien tulkintojen tekemisen esimerkiksi piilottamalla vinoumat.

3.4.2 LIME

LIME on vuonna 2016 esitetty algoritmi, jonka avulla voidaan selittää minkä tahansa luokittelija- tai regressiomallin approksimoimalla sitä paikallisesti, jolla tarkoitetaan, että selitetään yksittäistä ennustetta koko mallin sijaan. LIME-menetelmässä muuttujien tärkeyttä arvioidaan muuttamalla todellisia datanäytteitä ja tarkastelemalla koneoppimismallin tulosten muutosta muuttuneiden tapausten vuoksi ja rakentamalla siitä yksinkertainen paikallinen malli, joka approksimoi alkuperäisen mallin käyttäytymistä. (Linardatos et al. 2021; Moradi & Samwald, 2021; Ribeiro et al., 2016) Mittelstadt et al. (2019) mukaan LIME-menetelmä binarisoi ongelman. Sen sijaan, että menetelmä yrittäisi sovittaa lineaarista luokittelijaa suurelle arvoalueelle, voidaan jokainen muuttuja ajatella binäärisesti niin, että se voidaan kytkeä joko päälle tai pois. Tällöin on mahdollista vastata kysymykseen, että mikä kyseisen muuttujan vaikutus luokittelijan antamaan tulokseen. Tämä ei kuitenkaan vastaa siihen, että vaikutus verrattuna mihin. Jos kyseessä on strukturoimaton data, voidaan sitä verrata perustasoon. Strukturoidun datan osalta tämä on ongelmallisempaa. Miten esimerkiksi palkan merkitystä lainapäätöksen kannalta voidaan arvioida, jos luokittelija voi vain arvioida henkilöitä, joilla on riittävä suuri palkka. Sitä voitaisiin verrata toiseen riittävän suureen palkkaan, mutta kelvollisen palkan valitseminen on ongelma. (Mittelstadt et al., 2019)

Jos siis unohdetaan itse opetusdata ja kuvitellaan, että käytössä on vain niin sanottu mustan laatikon koneoppimismalli, johon syötetään dataa ja josta malli antaa ennusteen ja tavoitteena on ymmärtää, miksi malli päätyi kyseisen ennusteeseen. LIME käytännössä ottaen siis testaa, mitä ennusteille tapahtuu, kun annetaan muunneltua dataa koneoppimismalliin, jonka jälkeen LIME luo uuden datajoukon, joka koostuu muutetuista datanäytteistä ja sitä vastaavista mustan laatikon mallin ennusteista. Tätä uutta datajoukkoa LIME käyttää uuden tulkittavissa olevan mallin luomiseen. Käytettävä data voi olla tyypiltänsä joko teksti- tai taulukkomuotoista tai kuvia ja LIME onkin yksi harvoja menetelmiä, joka toimii näiden kolmen datatyypin kanssa. LIME-selitteitä on kuitenkin SHAP-menetelmän kaltaisesti mahdollista manipuloida muun muassa piilottamalla vinoumat, joka voi heikentää luottamusta LIME:n luomia selitteitä kohtaan. (Molnar, 2022)

3.5 Selittämisen haasteet

Arrieta et al. (2020) mainitsevat selitettävän tekoälyn haasteiksi muun muassa kompromissin tulkittavuuden ja tarkkuuden välillä, käsitteiden ja mittarien ymmärtämisen sekä suuntaviivojen luomisen tulkittavien tekoälymallien varmistamiseksi, selitettävän syväoppimisen saavuttamisen sekä tekoälyturvallisuuteen liittyvät selitteet. Arrieta et al. tutkimuksen mukaan kysymys tulkittavuudesta ja tarkkuudesta sisältää paljon myyttejä ja väärinymmärrystä. Esimerkiksi väite, että monimutkaisemmat mallit eivät välttämättä ole luonnostaan tarkempia, on virheellinen tapauksissa, joissa data on hyvin strukturoitu ja laadukasta. Mikä pitää paikkansa, on se, että monimutkaisilla malleilla on paljon enemmän joustavuutta kuin yksinkertaisilla vastineilla, mikä mahdollistaa monimutkaisempien toimintojen approksimoinnin. Tarkkuuden ja tulkittavuuden välinen kompromissi voidaan havaita tällaisissa tilanteissa, joissa datan strukturoimattomuus sekä lisätty monimutkaisuus heikentää tarkkuutta. Koneoppimismallille tehdyt selitykset pitäisi tehdä riittävän suuripiirteisiksi ja likimääräisiksi huomioiden kohdeyleisön vaatimukset, mutta varmistaen, etteivät selitteet yksinkertaista mallia liikaa. Esimerkiksi päätöspuut ja regressiomallit ovat tulkittavuudeltaan hyviä, mutta tarkkuudeltansa heikompia kuin esimerkiksi neuroverkot ja tukivektorikoneet, jotka ovat tarkempia, mutta heikommin tulkittavissa. (Arrieta et al., 2020)

Jotta selitettävän tekoälyn tutkimusala kehittyisi, olisi tärkeää luoda yhteinen perusta, jolle yhteisö voi luoda uusia tekniikoita ja menetelmiä. Siksi tulisi määrittää yhtenäinen selitettävyyden käsite, joka välittää alalla ilmaistuja tarpeita ja tarjoaa yhteistä pohjaa jokaiselle XAI-järjestelmälle. Toinen keskeinen haaste on oikeiden mittareiden ymmärtäminen. Mittareiden tulisi mahdollistaa vertailu siitä, kuinka hyvin malli sopii selitettävyyden määritelmään ja perinteisten mittareiden, kuten tarkkuuden ja herkkyyden, tulisi ilmaista kuinka hyvin malli toimii tietyllä selitettävyyden osa-alueella. Yleisesti ottaen XAI-mittauksissa tulee arvioida selitteiden hyvyyttä, hyödyllisyyttä ja tyytyväisyyttä sekä selittämisen vaikutusta mallin suorituskykyyn. Kvantitatiivisempia, yleisiä XAI-mittareita tarvitaan jatkossa enemmän tukemaan olemassa olevia mittausmenetelmiä ja työkaluja. Haasteena on lisäksi arvioida XAI-menetelmiä tosielämän asetuksilla. (Arrieta et al., 2020; Das & Rad, 2020)

Vaikka selitettävä tekoäly kehittyy jatkuvasti, varsinkin neuroverkot vaativat vielä monien haasteiden ylittämistä ennen kuin niitä voidaan selittää. Selitettävä tekoäly on vielä melko nuori ala, joten ympäröivästä sanastosta ja määritelmistä ei ole yksimielisyyttä ja esimerkiksi termeillä kuten muuttujien tärkeys ja merkityksellisyys viitataan kirjallisuudessa usein samaan käsitteeseen. Tekoälyn selitettävyyden haasteena on myös tarjota selitykset, jotka ovat yhteiskunnan, poliittisten päättäjien ja koko lain saatavilla.

Erityisesti ei-teknistä asiantuntemusta vaativien selitysten välittäminen on hyvin tärkeää sekä epäselvyyksien välttämiseksi että sosiaalisen oikeuden kehittämiseksi. (Arrieta et al., 2020)

4. ANALYYSI

Tässä luvussa käydään läpi työn analyysiprosessia. Aluksi käydään läpi yleisesti käytettyä dataa ja sen keräystä. Tämän jälkeen käydään läpi tarkemmin työssä käytettyä Agile CRISP-DM -prosessimallia.

4.1 Datan keräys

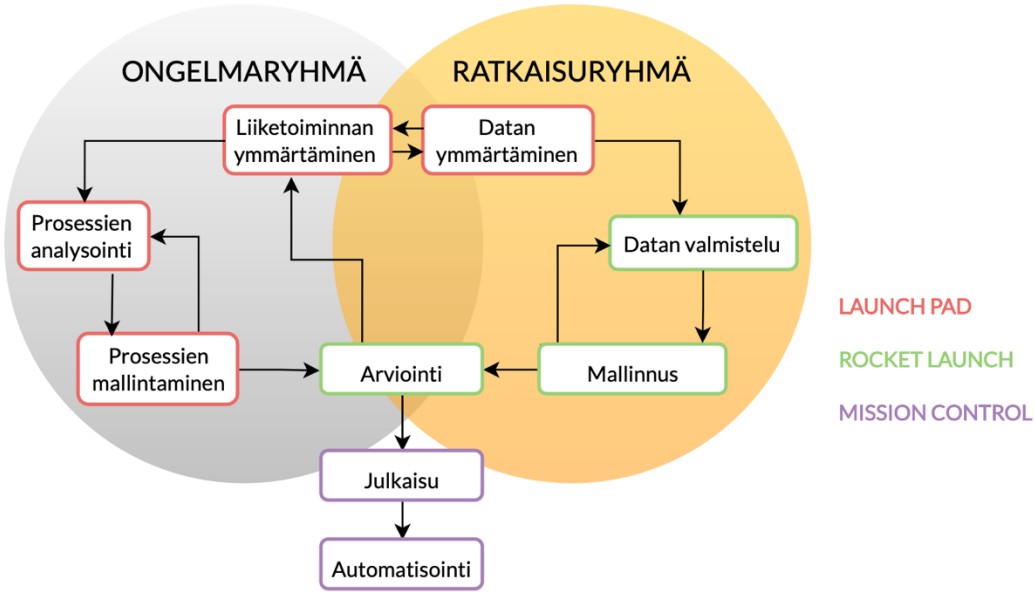
Työssä käytetyn koneoppimismallin pohjana toimii AIS-järjestelmästä (*eng. Automatic Identification System*) saatava historiadata. AIS on Kansainvälisen merenkulkujärjestön luoma järjestelmä alusten tunnistamiseen ja paikantamiseen, joka pyrkii siten estämään esimerkiksi alusten yhteentörmäyksiä. AIS-data sisältää tiedot muun muassa koordinaateista, nopeudesta ja suunnasta vallitsevalla ajanhetkellä sekä yleistietoa aluksesta kuten aluksen tunnuksen, alustyypin ja jääluokan. Työssä käytettiin Hakolan (2020) luomaa AIS-datasettiä eksploratiiviseen analyysiin sekä dataa haettiin suoraan myös Väyläviraston tietokannoista. Koska AIS-data ei anna tietoa jää-, eikä tuulitilanteesta, ja ovat merkittävässä roolissa talvimerenkulussa, koneoppimismallia varten tuli kerätä myös tuuli- ja jäädataa. Tuulidataa kerättiin Ilmatieteenlaitoksen (n.d.) säähavaintojen latauspalvelusta Kemin ja Kalajoen mittauspisteiltä. Jäädataa kerättiin Ruotsin Ilmatieteenlaitoksen tarjoamasta rajapinnasta, josta saa ladattua päiväkohtaisen jääkartan. AIS-, tuuli- sekä jäädatat yhdistämällä saatiin kattava datasetti koneoppimismallin opetusta varten. Koska tämän työn pääpaino kohdistuu koneoppimismallin selittämiseen, eikä itse datan mallintamiseen tai koneoppimismallin kehittämiseen, ei niitä käsitellä tässä luvussa. Työn lukemisen kannalta lukijan on kuitenkin hyvä ymmärtää yleisellä tasolla mitä dataa on kerätty.

4.2 Agile CRISP-DM

Työssä sovellettiin IBM:n (2016) kehittämään CRISP-DM-prosessimalliin pohjautuvaa ketterää Agile CRISP-DM -menetelmää (Kuva 5), joka on Houston Analyticsin (n.d.) kehittämä malli. Malli sopii hyvin tekoälyratkaisujen kehittämiseen ja sen ydinprosessi on sama kuin CRISP-DM-mallissa.

Malli koostuu kuudesta eri vaiheesta, joita ovat liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan valmistelu, mallinnus, arviointi ja toimeenpano sekä näiden lisäksi prosessin mallinnus ja analysointi. Agile CRISP-DM -menetelmässä prosessi jae-

taan kahdelle työryhmälle: ongelma- ja ratkaisuryhmälle. Ongelmaryhmä muodostetaan työstämään liiketoiminnallisia ongelmia ja ratkaisuryhmä puolestaan valmistamaan teknisiä ratkaisuja, joita ohjaavat toivotut liiketoiminnalliset tulokset. (Houston Analytics n.d.)



Kuva 5: Agile CRISP-DM (Mukaillen Houston Analytics n.d.)

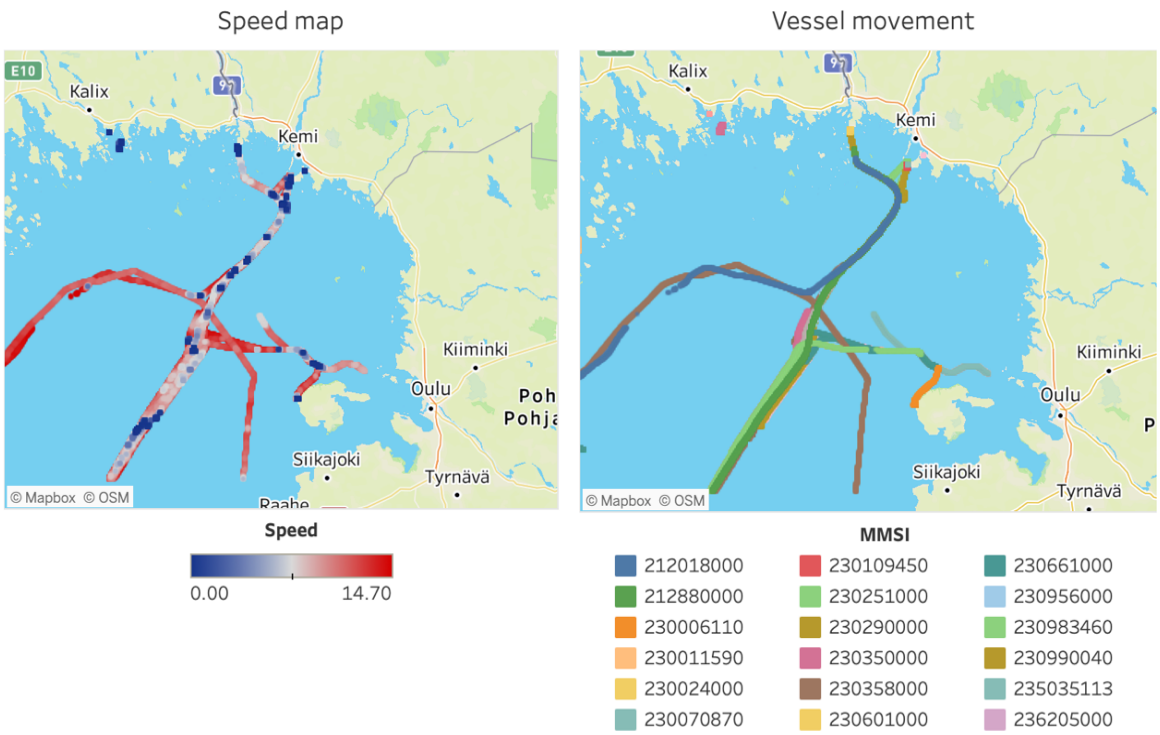
Agile CRISP-DM -menetelmässä prosessi jaetaan lisäksi kolmeen eri vaiheeseen, joiden englanninkieliset nimitykset ovat *Launch pad*, *Rocket launch* sekä *Mission control*. Ensimmäinen vaihe koostuu liiketoiminnan ja datan ymmärtämisestä sekä prosessien mallintamisesta ja analysoinnista. Tässä vaiheessa määritellään ja priorisoidaan tarpeet, luodaan ymmärrys nykyisistä prosesseista ja kerätään sekä määrällistä, että laadullista dataa. Toisessa vaiheessa valmistellaan ja mallinnetaan dataa sekä arvioidaan mallin toimivuutta. Jos malli ei täytä liiketoiminnallisia tarpeita, palataan ensimmäiseen vaiheeseen, muuten siirrytään viimeiseen vaiheeseen, jossa malli toimeenpannaan ja automatisoidaan. (Houston Analytics n.d.)

Tässä työssä ei ole tarkoitus luoda toimeenpantavaa tekoälymallia, vaan tutkimusluonteen vuoksi tutkia mahdollisia toteuttamiskelpoisia ratkaisuja, joten tutkimusprosessi on keskittynyt Agile CRISP-DM -mallin ensimmäiseen ja toiseen vaiheeseen sekä iteraatioihin näiden välillä. Palautetta kerättiin erillisissä verkkokokouksissa talvimerenkulun asiantuntijoiden kanssa ja saadun palautteen avulla tekoälymallia sekä selitettä kehitettiin iteratiivisesti. Lisäksi aineistoa kerättiin kirjallisten kysymysten avulla. Prosessissa oli mukana domain-asiantuntijoita eri aihepiireistä ja Agile CRISP-DM -menetelmän

mukaisesti työssä oli mukana asiantuntijoita niin ongelmatyöryhmästä, jotka toivat liiketoiminnallista näkökulmaa, kuin myös ratkaisutyöryhmästä, jotka kehittivät koneoppimismallia eli toimivat käytännön toteutuksen parissa. Seuraavissa alaluvuissa käydään läpi tutkimusprosessin iteraatioita ja miten koneoppimisratkaisuun on päästy ja miten selitettävyys on huomioitu.

4.2.1 Iteraatio 1

Agile CRISP-DM -menetelmän mukaisesti työ lähti liikkeelle datan ja liiketoiminnan ymmärtämisellä. Talvimerenkulun viranomaisten kanssa käydyistä verkkokokouksista muodostui ymmärrys talvimerenkulun toiminnasta sekä käytetyn datasetin muuttujista. Itse dataa lähdettiin aluksi tutkimaan visualisoimalla AIS-dataa Tableau-visualisointityökalulla. Kuvassa 6 on piirretty auki Hakolan (2020) keräämää AIS-dataa, jolla pyrittiin saamaan yleistä ymmärrystä datasta ja sen paikkansapitävyydestä. Kuvassa 6 on rajattu alue Perämereen aikaväliltä 17.3.–18.3. vuodelta 2018 eli yhden päivän ajalta. Vasemmanpuoleisesta nähdään esimerkiksi alueita, joissa alus pysähtyy tai nopeus putoaa lähelle nollaa, jolloin näillä alueilla alus on todennäköisesti juuttunut jääkenttään. Oikeanpuoleisesta visualisoinnista puolestaan näkee yleisesti eri alusten kulkemat reitit ja mahdollistaa tietyn aluksen valitsemisen tarkempaa tarkastelua varten.



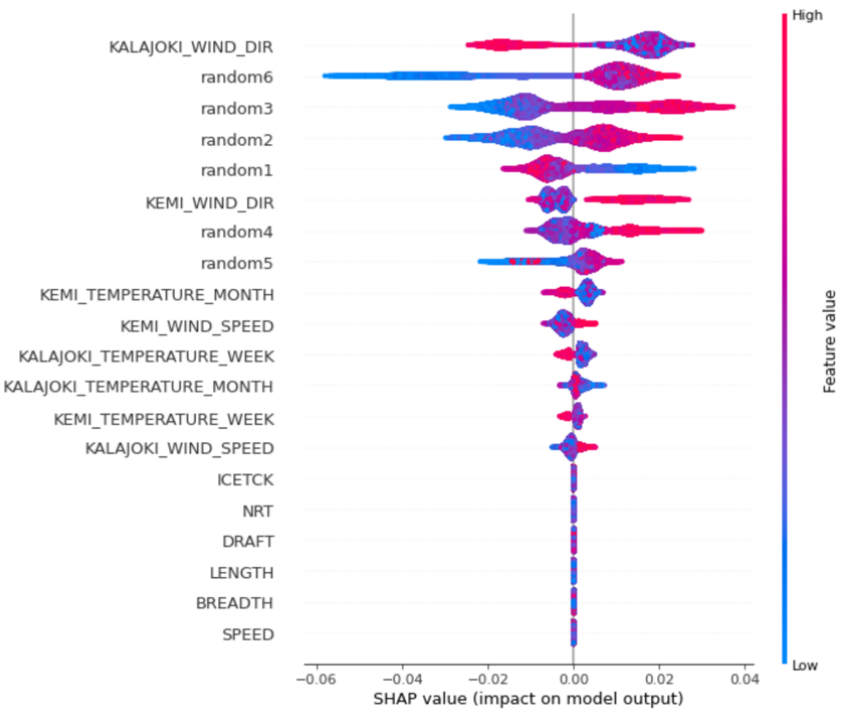
Kuva 6: Kuvakaappaus Tableau-visualisoinnista

Vaikkei tässä vaiheessa voida puhua vielä tekoälyn selitettävyydestä, on kuitenkin itse datan selittäminen keskeisessä roolissa, johon kuvaileva analytiikka on tehokas keino ja siten CRISP-DM-prosessimallin mukaisesti saavutetaan ymmärrystä datasta.

Datan eksploratiivisen analysoinnin jälkeen haasteena oli sopivan jäädatan löytäminen, jota ensimmäisessä iteraatiossa lähdettiin ratkomaan satelliittikuvien avulla. Tämä osoittautui hyvin haastavaksi useasta eri syystä. Satelliittikuvien lataus sekä käsittely koneoppimismalliin sopivaksi oli hyvin työlästä ja syvällistä osaamista vaativaa sekä suuri määrä satelliittikuvia vaati lisäksi paljon laskentatehoa. Talvimerenkulun asiantuntijoilta saadun palautteen mukaan satelliittikuvat ovat myös liian epätarkkoja, sillä satelliittikuvat otetaan kerran vuorokaudessa, mutta jäätilanne kulkuväylillä voi muuttua minuuteissa. Tämän vuoksi satelliittikuvien käyttö jätettiin heti pois jo heti prosessin alussa. Jään ollessa luonnollisesti keskeisessä roolissa talvimerenkulussa, päätettiin mallin kehittämisessä käyttää Ruotsin ilmatieteenlaitokselta saatavaa hilamuotoista jäädataa. Vaikka hiladata on yhtä epätarkkaa kuin satelliittikuvista saatava jäädata, oli sen työstäminen koneoppimismalliin sopivaksi helpompaa ja malliin saatiin jäätä edes jollain tasolla kuvaava muuttuja. Näin voitiin luoda alustava koneoppimismalli nopeuden ennustamiseen käyttäen Random forest -regressioalgoritmia. Itse koneoppimismallin luominen on tämän työn skaalan ulkopuolella, joten sen kehitysprosessia ei käydä tarkemmin läpi tässä työssä. Ensimmäisessä iteraatiossa oli kyse lähinnä prototyyppimallin luomisesta, joten selitettä ei lähdetty vielä tässä vaiheessa tekemään, vaan lähdettiin kehittämään mallia keskustelujen pohjalta seuraavassa iteraatiossa.

4.2.2 Iteraatio 2

Toisessa iteraatiossa koneoppimismallia lähestyttiin Pythonin SHAP-kirjaston avulla, jonka avulla mallia voitiin selittää ja siten eri muuttujien vaikutusta mallin antamiin tuloksiin oli mahdollista tarkastella syvällisemmin ja kehittäjille siten hyödyllistä parantamista varten. SHAP-selitteestä (Kuva 7) huomattiin, että varsinkin tuulen suunnalla oli suuri vaikutus mallin ennusteeseen. Toisaalta kuuden satunnaisen ”random”-muuttujan vaikutus koneoppimismalliin oli SHAP-selitteen mukaan suurta, josta voitiin päätellä, ettei mikään muuttuja vaikuttanut malliin merkittävästi, koska selitteen mukaan täysin satunnaisesti luoduilla muuttujilla on yhtä suuri vaikutus malliin kuin esimerkiksi tuulidatalla.



Kuva 7: SHAP-selite

SHAP-selitteen sai luotua Pythonilla hyvin yksinkertaisesti muutamalla rivillä seuraavasti:

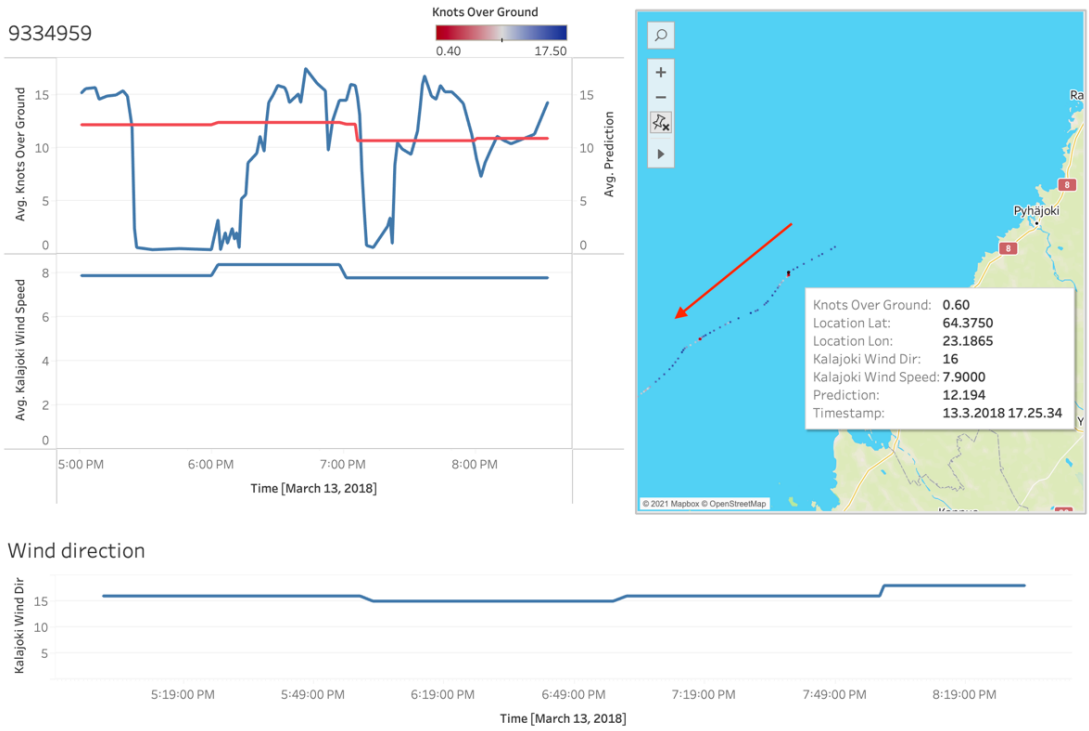
```
import shap
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(x_test)
shap.summary_plot(shap_values, x_test)
```

SHAP-selitteen lisäksi toisessa iteraatiossa kysyttiin myös suoraan talvimerenkulun asiantuntijalta aluksia, jotka olisivat sopivia esimerkkejä havainnollistamaan tyypillistä alusten kulkua Perämerellä ja siten hyviä aluksia koneoppimismallille.

Tyyppi	**IMO**	**Aluksen nimi**
Ro-ro-alus	9334959	Tavastland
Konttialus	9483669	X Press Elbe
Rahtialus	9142631	Ice Star
Rahtialus	9552032	Reymar
Säiliöalus	9390305	Stoc Marcia

Taulu 4: Valitut alukset

Valitut alukset ovat tyypeiltään Ro-ro-, kontti-, rahti- sekä säiliöaluksia ja taulussa 3 esitetyt alukset ovat tyypillisiä Pohjanlahdella kulkevia aluksia, jotka kulkevat säännöllisesti kyseisillä reiteillä. Jäänmurtajia ei valittu, sillä niiden kulkeminen jäissä on hyvin erilaista kuin esimerkiksi rahtialuksella avustuksen antamisesta johtuen, jolloin jäänmurtajat olisivat heikentäneet koneoppimismallin tarkkuutta ja väärentäneet laivan kulkua. Valittujen alusten nopeutta piirrettiin auki Tableaulla ja verrattiin malliin antamiin nopeusennusteisiin, jonka avulla nähtiin selkeästi kuinka tarkasti malli toimii. Kuvassa 8 on piirretty auki Tavastland-rahtilaivan kulkua ja ennustetta Perämerellä maaliskuun 13. päivä vuonna 2018, jolloin nopeudessa oli havaittavissa kaksi äkillistä putoamista. Rahtilaiva matkusti Kemistä etelään päin ja Kalajoen edustalla nopeus putosi äkillisesti hyvin lähelle nollaa yli 15 solmun matkavauhdista. Hidastumisen hetkellä Kalajoen mittauspisteellä tuulen nopeus oli 7,9 m/s ja tuulen suunta 16 astetta eli pohjoistuulta. Talvimerenkulun asiantuntijan mukaan rannikolla tämän suuruinen tuuli vastaa avomerellä yli 10 m/s, jolloin kyse on nopeudesta, joka voi vaikuttaa laivan kulkuun rännissä.



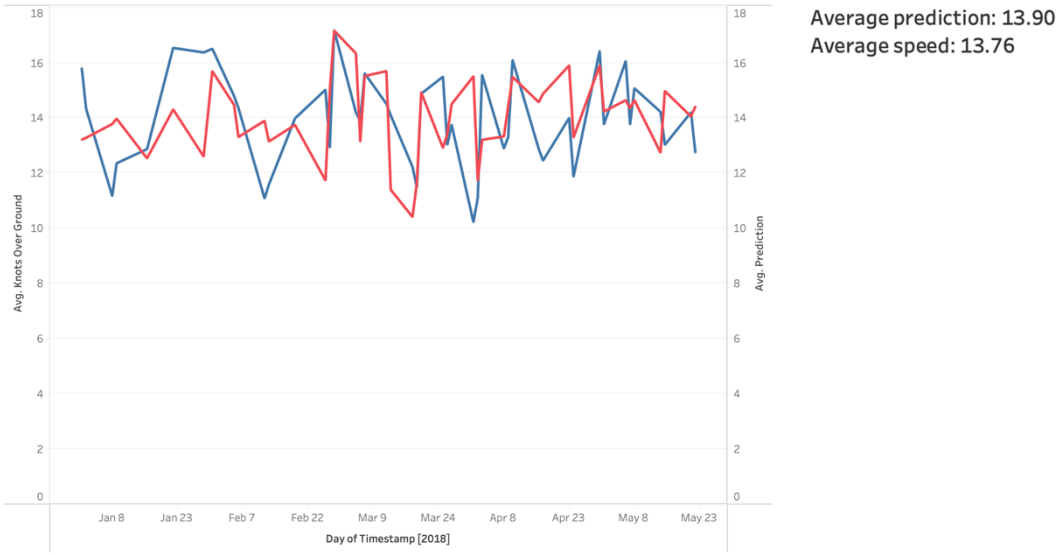
Kuva 8: Tavastland-aluksen kulkeminen

Laivan kulkiessa lounaaseen, osuu pohjoistuuli laivan oikeaan kylkeen, joka voi voimakkaalla tuulella liikuttaa laivaa rännin vasempaan reunaan ja siten voi aiheuttaa laivan hidastumista tai jopa pysähtymisen.

Kuvan 8 vasemman yläkulman kuvaajassa on sinisellä esitetty aluksen toteutuneen nopeuden keskiarvo solmuissa ja punaisella koneoppimismallin ennustaman nopeuden

keskiarvo. Tämän alapuolella on Kalajoen mittauspisteeltä mitattu tuulidata. Huomataan, että koneoppimismallin luoma ennuste seuraa suoraan verrannollisesti tuulen nopeuden muutosta eli tuulen nopeuden kasvaessa siis myös laivan nopeus kasvaisi. Talvimerenkulun asiantuntijan mukaan tuulen nopeutta tärkeämpi muuttuja on tuulen suunta. Kuvassa esitetyssä tapauksessa suuremmalla tuulennopeudella voi olla nopeuttavaa vaikutusta, sillä se osuu viistosti laivan kylkeen eikä kohtisuoraan. Jos tuulen suunta olisi koillistuulta eli osuisi laivan perään, olisi tuulen nopeuttava vaikutus suurimmillaan ja keulaan osuessaan luonnollisesti päinvastainen.

Muutaman tunnin lyhyttä ennustetta tärkeämpää onkin tutkia pidemmän aikavälin keskiarvoennustetta, jolloin esimerkiksi saapumisajan ennustaminen olisi tarkempaa, kun malli ei huomioisi yksittäisiä pysähdyksiä. Kuvasta 9 nähdään, että Tavastland-aluksen pitkän aikavälin ennuste seuraa toteutuneen nopeuden keskiarvoa hyvin tarkasti, sillä ennusteen ja toteutuneen nopeuden ero on vain kymmenyksiä.

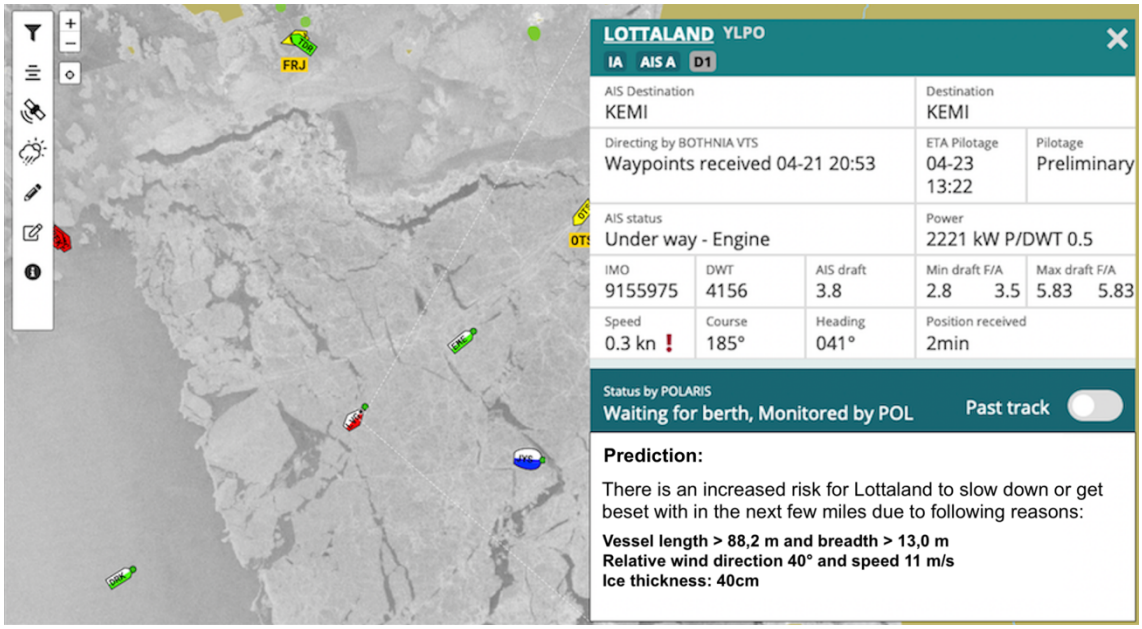


Kuva 9: Pitkän aikavälin ennuste

4.2.3 Iteraatio 3

Kolmannessa iteraatiossa mallia pyrittiin vielä kehittämään tuulen osalta ja uudeksi muuttujaksi saatiin suhteellinen tuulen suunta. Kuten edellisessä iteraatiossa huomattiin, tuulen suhteellinen suunta on mallille parempi muuttuja kuin absoluuttinen tuulen suunta, sillä tuuli voi vaikuttaa nopeuteen joko kasvattaen sitä tai hidastaen riippuen suunnan suhteesta alukseen. Tämä saatiin laskettua AIS-datasta saatavan aluksen suunnan ja tuulen suunnan erotuksena, jonka jälkeen suhteellinen tuulen suunta vielä jaettiin 8 eri sektoriin eli 45 asteen sektoreihin. Selitteiden ansiosta tuuleen ja varsinkin

sen suuntaan kiinnitettiin enemmän huomiota, joka olisi ilman selitteitä jäänyt todennäköisesti huomioimatta. Voidaan siis todeta, että selitettävä tekoäly auttoi mallin kehitysprosessissa. Kuitenkin kehittäjät olivat vain yksi sidosryhmä ja talvimerenkulusta puhuttaessa tärkeää on, että varsinainen loppukäyttäjä eli se, jonka käyttöön tekoälymallia luodaan, kuten jäänmurtajan kapteeni, saa helposti tulkittavan selitteen mallista. Selite voisi olla esimerkiksi upotettuna IBNet-järjestelmään, joka on hajautettu järjestelmä jäänmurtajien seurantaan ja ohjaamiseen talvimerenkulussa. Kuvan 10 IBNet:n käyttöliittymän oikeaan alakulmaan on lisätty hahmotelma mahdollisesta selitteestä, jonka voi luoda SHAP-selitteen pohjalta.



Kuva 10: Hahmotelma loppukäyttäjän selitteestä

Talvimerenkulun asiantuntijan mukaan tällainen tieto on kuitenkin nyt ja vielä lähivuosinakin erittäin epäluotettavaa johtuen muun muassa jääkentän dynamiikasta, ajantasaisen jäätiedon puutteesta, alusten teknisen jäissäkulkukyvyn todellisen ennusteen tarkkuudesta sekä alusta ohjaavan henkilön vaikutuksesta. Aluksen ympärillä muutamien kymmenien metrien tarkkuudella vallitseva jäätilanne on kuitenkin tärkein muuttuja, josta ei ole vielä mahdollista saada tarkkaa tietoa. Saapumisajan ennustamisen talvimerenkulun asiantuntija näkee kuitenkin mahdollisena ja hyödyllisenä kehityspolkuna, sillä se perustuu keskiarvojen kautta syntyvään oletukseen, jonka tarkkuutta voitaisiin koneoppimisen kautta kehittää eri vaikuttavien tekijöiden keskinäisten painokertoimien merkitystä tarkentamalla. Tässä työssä ei kuitenkaan lähtökohtanaan ollut käyttää selitteiden pohjana tuotantovalmista koneoppimismallia, vaan tutkia miten selitettävyys voisi tulevaisuudessa näkyä talvimerenkulussa. Kuvassa 10 esitetty selite on siten vain

esimerkki mahdollisesta selitteestä. Tärkeintä on, että selite on loppukäyttäjälle selkeä ja helposti ymmärrettävissä, oli kyseessä sitten tekstipohjainen selite tai visuaalinen mittari. Agile CRISP-DM -mallin viimeiset vaiheet eli julkaisu ja automatisointi jätettiin luonnollisesti tästä työstä pois.

4.3 Selitettävyyden tasot talvimerenkulussa

Kuten aikaisemmin on mainittu, selitettävällä tekoälyllä on olemassa eri sidosryhmätasoja, joille selitettävyys voi tarkoittaa eri asioita. Tässä työssä sidosryhmiä on käyty kehittäjien, teoreetikoiden ja toimialaosaajien, eetikoiden sekä käyttäjien näkökulmasta. Edellä esitetyissä iteraatioissa asiaa on käsitelty enemmän kehittäjien, toimialaosaajien ja loppukäyttäjien näkökulmasta, mutta myös eetikoiden rooli talvimerenkulun kontekstissa voidaan nähdä tärkeänä, sillä he keskittyvät tekoälyratkaisujen oikeudenmukaisuuteen, vastuullisuuteen sekä läpinäkyvyyteen. Kukin näistä on tärkeitä osa-alueita tulevaisuudessa säännöstelyn tiukentuessa, jos talvimerenkulun tekoälyratkaisua aloitetaan kehittämään. Siten myös eetikoille kohdistetut selitteet tulisivat olla vähemmän teknisiä ja helpommin ymmärrettäviä, jotta eettiset asiat huomioidaan paremmin kehittäessä tekoälyratkaisuja.

Talvimerenkulussa eettiseen tason kysymykset voivat koskea esimerkiksi jäänmurtajien kauppa-aluksille antamaa avustusta. Jos kehitetty malli pyrkisi esimerkiksi optimoimaan jäänmurtajien tarjoamaa avustusta, tulisi miettiä, että mihin oikea avustusjärjestys perustuu. Kysymykseen, että mikä on oikea, voi eetikon tuoma näkökanta antaa vastauksen. Perustuisiko oikea avustusjärjestys esimerkiksi kauppa-aluksen lastin arvoon, huoltovarmuuden varmistamiseen tai laivojen määrään. Pelkän tekoälymallin kehittäjän on mahdotonta luoda toimivaa ratkaisua tällaisissa tapauksissa ilman eetikon tuomaa näkökulmaa. Eetikoiden lisäksi toimialaosaajien tuoma käytännön osaaminen mahdollistaa mallin ja sitä kuvaavien selitteiden paremman ymmärtämisen ja tarvittaessa siten mallin jatkuvan kehittämisen.

Talvimerenkulkuun liittyy Väyläviraston lisäksi monia eri sidosryhmiä, kuten satamat ja niiden operaattorit, teollisuus sekä jäänmurtajien ja kauppa-alusten kapteenit. Tässä työssä käydyn aluksen nopeuden ennustemallin lisäksi mahdollisuuksia erilaisten tekoälyratkaisujen kehittämiseksi on siis lukuisia. Siten myös kaikkien selitettävän tekoälyn sidosryhmien hyödyntäminen mallin kehittämisessä on tärkeää ja luo pohjan kestävälle ratkaisulle.

5. TULOKSET JA POHDINTA

Työn tavoitteena oli tutkia selitettävää tekoälyä talvimerenkulussa. Johdannossa annettiin yleiskuva talvimerenkulusta ja perehdyttiin tutkimusmenetelmiin hyödyntäen Saunders et al. (2009) sipulimallia. Tämän jälkeen syvennyttiin työn teoriaosuuteen (luvut 2 & 3), jonka jälkeen käytiin läpi työn analyysia. Tässä luvussa esitetään pohdintaa työstä sekä käydään läpi työn tuloksia tiivistetysti. Luku aloitetaan yhteenvedolla, jossa käydään läpi tutkimuskysymykset sekä yleisesti pohdintaa tuloksista ja löydöksistä. Tämän jälkeen arvioidaan tuloksia ja lopuksi pohdintaan jatkotutkimusmahdollisuuksia.

5.1 Yhteenveto

Tutkimuksen tavoitteena oli löytää selitettävän tekoälyn tekijöitä talvimerenkulussa, joita lähdettiin tutkimaan kolmen tutkimuskysymyksen kautta. Näitä olivat: *Miksi tekoälyn selitettävyyttä tarvitaan talvimerenkulussa?, miten selitettävä tekoäly auttaa koneoppimismallin kehittämisessä? sekä kuinka hyvin tekoälyselitteitä voidaan ymmärtää eri sidosryhmien näkökulmasta?* Näihin saatiin kirjallisuuden sekä analyysissa esitetyn prosessin avulla vastattua monipuolisesti. Seuraavaksi käydään työhön valittuja tutkimuskysymyksiä tarkemmin läpi.

Miksi tekoälyn selitettävyyttä tarvitaan talvimerenkulussa?

Talvimerenkulussa jää on keskeinen muuttuja ja saatava jäädata on nykyisellään ja todennäköisesti vielä tulevina vuosinakin hyvin epätarkkaa. Talvimerenkulun asiantuntijan mukaan avomerellä jääkenttä elää ja muuttaa olomuotoansa jatkuvasti sen olemassaolonsa aikana. Tässä työssä kuitenkin keskityttiin selitettävyyteen ja oletettiin tulevaisuuden tilanne, jossa riittävän luotettava tekoälymalli olisi jo olemassa.

Selitettävän tekoälyn tarpeella todettiin olevan monia erilaisia perusteita. Luottamuksen, läpinäkyvyyden ja ymmärryksen luonti on keskeisessä roolissa talvimerenkulun tekoälyratkaisuissa. Jos aluksen kuljettamisessa hyödynnetään tekoälyä, on luottaminen ja tehtyjen päätösten läpinäkyvyys ensiarvoisen tärkeää. Tulevaisuudessa yhä enenemässä määrin erilaiset lait ja regulaatiot tulevat vaatimaan läpinäkyvyyttä tekoälyratkaisuilta ja selitettävyys on keino vastata näiden vaatimuksiin. Esimerkiksi Viljasen (2021) mukaan EU:n ehdotus tekoälyteknologioita sääntelevästä asetuksesta koskee korkean riskin tekoälyjärjestelmiä, johon myös muun muassa meriturvallisuuslaitteet, ja siten tal-

vimerenkulku, kuuluvat. Täten jo heti kehitysvaiheesta alkaen selitettävyyden huomioiminen aina loppukäyttäjälle saakka luo kyvyn vastata säännösten vaatimiin kysymyksiin. Edellä mainittuun liittyen myös sosiaalisen vastuun, oikeudenmukaisuuden ja riskien huomioiminen on selitettävän tekoälyn etu. Kun kyse on laivaliikenteestä, voi virheiden hinta olla rahallisesti hyvin suuria tai pahimmillaan jopa henkilövahinkoja aiheuttavia. Lisäksi talvimerenkulussa voi tulla eteen avustukseen liittyviä kysymyksiä, kuten että käyttääkö alus kaikkea moottoritehoansa polttoaineen säästämiseksi ja jää siksi kiinni, kun tietää, että avustus on ilmaista. Tekoälyratkaisu voisi tunnistaa tällaiset alukset ja hyvin tulkittavat selitteet voisivat mahdollistaa nopean reagoinnin.

Miten selitettävä tekoäly auttaa koneoppimismallin kehittämisessä?

Selitettävä tekoäly tulisi ajatella osana koko tekoälymallin kehittämisprosessia, eikä vain lopputuotteena, selitteenä, jossa mallin antamaa tulosta selitetään. Selitettävyyden huomioiminen heti prosessin alkuvaiheessa mahdollistaa vinoumien ja väärinkäsitysten minimoinnin tai esimerkiksi uusien muuttujien luomisen malliin. Tässä työssä käytiin läpi kaksi yleistä selitettävän tekoälyn menetelmää: SHAP ja LIME. Molemmat näistä ovat varsinkin kehitysvaiheessa tehokkaita työkaluja selittämään koko mallia tai yksittäistä ennustetta ja helposti käytettävissä Python-ohjelmointikielellä. Selitettävä tekoäly auttaa Agile CRISP-DM-prosessimallin arviointivaiheessa arvioimaan mallin tarkkuutta ja muuttujien vaikutusta ja tarvittaessa palaamaan prosessimallin alkuun esimerkiksi tutkimaan dataa ja käsittelemään sitä uudestaan. Esimerkiksi tässä työssä mallia arvioitaessa selitettävän tekoälyn avulla tuuleen osattiin kiinnittää tarkempaa huomiota, jonka ansiosta saatiin suhteellinen tuulen suunta uudeksi muuttujaksi.

Kehittäjätasolla selitettävyys antaa siis paremman mahdollisuuden ymmärtää mallia, sen muuttujia ja niiden riippuvuuksia sekä kehittäjien itse hyödyntää rationaalista ajattelua. SHAP-yhteenvetokuvaajasta nähdään muuttujien vaikutus malliin, jolloin tietyn muuttujan epäoletettua vaikutusta malliin voidaan tarkastella lähemmin ja kehittäjä pystyy kysymään tarkempaa lisätietoa muilta sidosryhmiltä, kuten toimialaosaajilta. Hong et al. (2020) mukaan kuitenkin kehittäjien suhdetta selitettävyyteen ja miten selitettävyys näkyy kehittäjien käytännöissä ja tarpeissa on tutkittu vielä suhteellisen vähän.

Kuinka hyvin tekoälyselitteitä voidaan ymmärtää eri sidosryhmien näkökulmasta

Eri sidosryhmille selitettävyys voi tarkoittaa eri asioita. Aina näin ei ole, mutta pääsääntöisesti voidaan olettaa tekoälymallin kehittäjän tarve selitettävyydelle olevan hyvin erilainen, kuin selitettävyys loppukäyttäjälle. Tässä työssä selitettävän tekoälyn sidosryhmiä käytiin läpi neljän eri sidosryhmän kautta, joita ovat kehittäjät, teoreetikot ja toimialaosaajat, eetikot sekä käyttäjät, joista jokaisessa on nähtävissä omat selitettävyyden ominaispiirteensä. Edellisessä alaluvussa käytiin läpi selitettävyyttä läpi kehittäjien ja mallin kehitysprosessin näkökulmasta. Teoreetikot ovat vahvasti sidoksissa kehittäjien kanssa ja voivat olla kehittäjien tavoin tekoälymallien luojia. Heidän tavoitteenansa on kehittää itse tekoälyä käytännön tekoälyjärjestelmien kehittämisen sijaan. Toimialaosaajat puolestaan ymmärtävät syvemmin toimialasta ja voivat auttaa varmistamaan, että selitteet täyttävät tarvittavat kriteerit, mutta tekoälymalleista heillä on harvemmin ymmärrystä. Heille selitteet voivat olla esimerkiksi kehittäjän luoma SHAP-selite, jota käydään yhdessä kehittäjän kanssa läpi molemman puoleisen tulkittavuuden takaamiseksi. Toisaalta toimialaosaaja voi myös olla mallin loppukäyttäjä, jolloin selite on erilainen.

Eetikoihin kuuluvat henkilöt, jotka keskittyvät muun muassa tekoälymallien oikeudenmukaisuuteen, luotettavuuteen sekä vastuullisuuteen. Eetikot voivat kuulua myös edellä mainittujen sidosryhmien tasolle, mutta selitteiden tulisi lähestyä mallia vähemmän tekniseltä kantilta. Koska eetikoiden tavoitteena on saada mallista eettisesti hyväksyttävä, tulee heidän panostansa hyödyntää kehitysvaiheessa. Loppukäyttäjät eroavat edellä esitetyistä, sillä heille selite on lopputuote, jonka tulee olla mahdollisimman selkeä ja helposti tulkittavissa, tekoälymallin ”jäävuoren huippu”. Se ei sisällä mallin syvällisiä arvoja, vaan selite voi olla yksinkertainen visualisointi tai tekstiselite, joka esimerkiksi varoittaa tulevasta vaarasta perusteluineen. Voidaan siis todeta, että selitettävä tekoäly koskee kaikkia prosessiin osallistuvia henkilöitä, vaikka sitä voidaan ymmärtää eri tavalla eri sidosryhmien kesken.

5.2 Tulosten arviointi

Tässä työssä käytetty aineisto on kerätty kirjallisuudesta, kvalitatiivisesta keskusteluaineistosta sekä kvantitatiivisesta datasta. Vaikka selitettävää tekoälyä ja talvimerenkulkuun liittyvää mallinnusta on tutkittu erikseen, oli näiden aihealueiden yhdistäminen haastavaa kirjallisuuden vähyyden vuoksi. Talvimerenkulun asiantuntijoiden kanssa käydyistä keskusteluista saatiin kuitenkin arvokasta tietoa talvimerenkuluun liittyen ja

kirjallisuuden avulla pystyttiin vastaamaan selitettävään tekoälyyn liittyviin kysymyksiin. Kvantitatiivista dataa työssä käytettiin selitteiden luomiseen tämän työn ulkopuolella kehitettyyn koneoppimismalliin pohjautuen.

Eräs haaste selitettävyyden arvioinnissa on, ettei selitettävyyttä voida absoluuttisesti mitata, toisin kuin selitteen pohjana toimivaa mallia, jossa esimerkiksi ennusteen tarkkuutta toteutuneeseen voidaan hyvin verrata. Käyttäjätasolla selitteiden käytön myötä saatava palaute mahdollistaa selitteen kehittämisen, jota ei tässä työssä luonnollisesti ollut mahdollista saada. Selitteiden pohjana ei myöskään ollut tuotantovalmista koneoppimismallia, joten selite oli hahmotelma, miltä se voisi tulevaisuudessa näyttää talvimerenkulussa. Työssä kuitenkin onnistuttiin luomaan SHAP-selite koneoppimismallin pohjalta, josta oli hyötyä koneoppimismallin kehittämisprosessissa. Voidaan myös arvioida, että tutkimuskysymyksiin onnistuttiin vastaamaan.

Teoriaosuuksiin pystyttiin vastaamaan kirjallisuuden sekä talvimerenkulun asiantuntijoiden kanssa käytyjen keskustelujen pohjalta monipuolisesti. Aluksi käytiin läpi tekoälyn ja laivaliikenteen teoriaa, josta siirryttiin käsittelemään selitettävän tekoälyn teoriaa. Kirjallisuuden osalta haasteena oli talvimerenkulkua ja selitettävää tekoälyä yhdistävän kirjallisuuden puute, jonka vuoksi kirjallisuudelle ei tehty erityistä suodatusta eikä sitä voitu verrata muihin alan julkaisuihin. Selitettävää tekoälyä koskevaa kirjallisuutta on paljon, mutta yhdistettynä pääosin täysin eri aihealueeseen kuin talvimerenkulkuun. Samoin talvimerenkulkuun liittyvää kirjallisuutta on, mutta ne keskittyvät itse mallintamiseen tai esimerkiksi ennusteiden luomiseen, mutta huomioimatta selitettävyyttä. Näitä yhdistävä ”lanka” saatiin keskusteluista talvimerenkulun asiantuntijoiden kanssa.

Työn tutkimuskysymyksiin pystyttiin vastaamaan pääosin kirjallisuuden avulla, joten analyysi oli siinä määrin enemmän kirjallisuutta tukeva kuin uutta teoreettista ja käytännöllistä kontribuutiota tuottava. Agile CRISP-DM-prosessimalli on tekoälyä ja sen selitettävyyttä luodessa monipuolinen vaihtoehto, mutta tämän työn kehyksessä liiankin monipuolinen, eikä kaikkia osa-alueita ollut mahdollista hyödyntää, jolloin työn analyysi oli lähempänä perinteisempää CRISP-DM-mallia. Kaiken kaikkiaan työssä saavutettiin kuitenkin halutut tulokset. Seuraavaksi pohditaan vielä, millaisia jatkotutkimusmahdollisuuksia aiheella on.

5.3 Jatkotutkimus

Kuten edellä on mainittu, selitettävä tekoäly ja talvimerenkulku ovat nykyisellään vielä hieman irrallaan olevia aihealueita, sillä jään luoman haasteen vuoksi tekoälyjärjestelmien kehittäminen ja käyttö talvimerenkulussa on vielä hyvin haastavaa. Tämän vuoksi

ei ole myöskään tekoälyä, jota selittää. Jatkotutkimuksen näkökulmasta olisi hyvä tutkia aihealueita erikseen ja siten luoda pohja selitettävyydelle. Tarkoittaen, miten esimerkiksi talvimerenkulkuun pystytään luomaan tarkempaa jäämallinnusta, joka mahdollistaisi tarkkojen ennustemallien luomisen ja siten niiden selittämisen tai puolestaan miten selitettävää tekoälyä voisi kehittää niin, että prosessi olisi yhtenäinen kaikkien sidosryhmätasojen kanssa. On vielä hyvin haastavaa puhua selitettävästä tekoälystä talvimerenkulussa ennen kuin pohja selitettävyydelle on kunnossa. Tässä työssä tutkittiin aihetta tulevaisuuden tilanteen näkökulmasta, jossa pohja olisi jo kunnossa ja selitettävää tekoälyä olisi mahdollista luoda. Yleisesti ottaen selitettävään tekoälyyn liittyen jatkotutkimusta voisi tehdä muun muassa siitä, miten selitettävyyden saisi osaksi tyypillistä koneoppimismallin kehitysprosessia ja miten siitä tulisi ”uusi normaali” tekoälystä puhuttaessa. Esimerkiksi jatkojalostamalla jo olemassa olevia prosessimallia tai kehittämällä uuden, joka huomioisi selitettävyyden paremmin.

LÄHTEET

Adadi, A & Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).

Alpaydin, E. 2014. Introduction to Machine Learning, MIT Press

Anttila, P. 1998. Tutkimisen taito ja tiedonhankinta. Saatavilla: <https://metodix.fi/2014/05/17/anttila-pirkko-tutkimisen-taito-ja-tiedon-hankinta/#7.1%20Tieteellisen%20päättelyn%20logiikat> (Viitattu 18.6.2021)

Arctia. 2016. Jäänmurron prosessikuvaus. Youtube-video. Julkaisija Arctia ltd. Saatavilla: <https://www.youtube.com/watch?v=BJJ9n-7Lkxc> (Viitattu 1.2.2022)

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.

Bhatt, U. & Xiang, A. % Sharma, S. & Weller, A & Taly, A. & Jia, Y. & Ghosh, J & Puri, R. & Moura, J. & Eckersley, P. 2020. Explainable Machine Learning in Deployment.

Bonaccorso, G. 2018. Mastering Machine Learning Algorithms: Expert Techniques to Implement Popular Machine Learning Algorithms and Fine-Tune Your Models.

Collis, J. & Hussey, R. 2003 Business Research: A Practical Guide for Undergraduate and Postgraduate Students (2nd edn). Basingstoke: Palgrave Macmillan. (Katso viittaustekniikka, Saunders oli viitannut tähän)

Confalonieri, R & Coba, L & Wagner, B & Besold, T. 2019. A historical perspective of explainable Artificial Intelligence.

Das, A. and Rad, P., 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey.

DeChant, J. 2019. How AI is Influencing the Shipping Industry Today.

Doran, D & Schultz, S & Besold, T. 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.

Fink, A. (2014) Conducting research literature reviews: from the Internet to paper. 4th ed. Thousand Oaks (Calif.): Sage.

Gerlins, J., Shollo, A. & Constantiou, I. 2021. Reviewing the Need for Explainable Artificial Intelligence (xAI)

Gosiewska, A & Biecek, P. 2020. Do not trust additive explanations.

Hakola, V. 2020. PREDICTING MARINE TRAFFIC IN THE ICE-COVERED BALTIC SEA. Saatavilla: <https://trepo.tuni.fi/bitstream/handle/10024/120219/Hakola-Ville.pdf?sequence=2&isAllowed=y>

Hakola, V. 2020. Saatavilla: VESSEL TRACKING (AIS), VESSEL METADATA AND DIRWAY DATASETS <https://ieee-dataport.org/open-access/vessel-tracking-ais-vessel-metadata-and-dirway-datasets>

Hastie, T. 2009. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition.

Hong, S. & Hullman, J. & Bertini, E. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs.

Houston Analytics. (n. d.) Crafting improved business operations through agile analytics projects. <https://www.houston-analytics.com/project-methodology> (Viitattu 8.7.2021)

Hurwitz, J. & Kirsch, D. 2018. Machine Learning For Dummies. Saatavilla: <https://www.ibm.com/downloads/cas/GB8ZMQZ3> (Viitattu 2.11.2021)

IBM. 2016. IBM SPSS Modeler CRISP-DM Guide. <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf> (Viitattu 8.7.2021)

IBM Cloud Education. 2020. Strong AI. <https://www.ibm.com/cloud/learn/strong-ai#toc-strong-ai--YaLcx8oG> (Viitattu 1.11.2021)

VTT. 2009. IBNet - jäänmurtajien koordinointi. Saatavilla: <https://projectsites.vtt.fi/sites/ibnet/www.vtt.fi/sites/ibnet.html> (Viitattu 1.4.2022)

Ilmatieteenlaitos, N.d. Havaintojen lataus. <https://www.ilmatieteenlaitos.fi/havaintojen-lataus> (Viitattu 1.4.2022)

Ilmatieteenlaitos. N.d. Jäätalvi Itämerellä <https://www.ilmatieteenlaitos.fi/jaatalvi-itamerella> (Viitattu 25.6.2021)

Jin, W., Fan, J., Gromala, D., Pasquier, P. and Hamarneh, G., 2021. EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A. and Baum, K., 2021. What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research.

Launis, R. 2020. Jakautunut Suomi, selitettävä tekoäly ja yksittäiset postinumerot. Saatavilla: <https://www.launis.net/post/jakautunut-suomi-selittävä-tekoäly-ja-yksittäiset-postinumerot> (Viitattu 5.8.2021)

Leong, L. 2019. Maritime Fairtrade. AI Revolution: 6 Steps To Prepare Your Business. Saatavilla: < https://maritimefairtrade.org/ai-revolution-6-steps-to-prepare-your-business> (Viitattu 29.12.2021)

Lehtola, V, Montewka, J, Goerlandt, F, Guinness, R & Lensu, M. 2019. Finding safe and efficient shipping routes in ice-covered waters: A framework and a model, Cold Regions Science and Technology

Liikenne- ja viestintäministeriö. 2014. Suomen meriliikennestrategia 2014–2022. Saatavilla: <https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/77909/Julkaisuja_9-2014.pdf?sequence=1> Viitattu: 20.10.2021

Lin, Y., Lee, W. & Celik, Z. 2020. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors.

Linardatos, P. & Papastefanopoulos, V. & Linardatos, P. 2021. Explainable ai: A review of machine learning interpretability methods.

Liu, B., 2021. " Weak AI" is Likely to Never Become" Strong AI", So What is its Greatest Value for us?

Lipton, Z. C. 2017. The Mythos of Model Interpretability.

Lundberg, S. & Lee, S. 2017. A unified approach to interpreting model predictions.

Lundberg, S. 2019. An introduction to explainable AI with Shapley values. < https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html> (Viitattu: 10.2.2022)

Lähdesmäki, T., Hurme, P., Koskimaa, R., Mikkola, L., Himberg, T., Menetelmäpolkuja humanisteille. Jyväskylän yliopisto, humanistinen tiedekunta. <http://www.jyu.fi/mehu>. (Viitattu 22.6.2021)

MacKinnon, S. & Weber, R. & Olindersson, F. & Lundh, Monica. 2020. Artificial Intelligence in Maritime Navigation: A Human Factors Perspective.

Mahendran, A. and Vedaldi, A. 2015. Understanding deep image representations by inverting them.

Mittelstadt, B., Russell, C. and Wachter, S. 2019. Explaining explanations in AI.

Montewka, J & Goerlandt, F & Kujala, P & Lensu, M. 2014. Towards probabilistic models for the prediction of a ship performance in dynamic ice.

Molnar, C. 2022. A Guide for Making Black Box Models Explainable. Saatavilla: <https://christophm.github.io/interpretable-ml-book/> Viitattu: 19.2.2022.

Moradi, M. & Samwald, M. 2021. Post-hoc explanation of black-box classifiers using confident itemsets. Expert systems with applications.

Ojala, L. & Kujala, P. & Solakivi, T. & Kiiski, T. & Lindeberg, M. & Kilpi, V. 2020. Merlog 2030 Merikuljetusten logistiikka ja ulkomaankaupan kilpailukyky.

Pietikäinen, M. & Silvén, O. 2021. Tekoälyn haasteet – Koneoppimisesta ja konenäöstä tunnetekoälyyn. 2. Painos.

Preece, A., Harbone, D., Braines, D., Tomsett, R. & Chakraborty, S. 2018. Stakeholders in Explainable AI. https://arxiv.org/abs/1810.00184

Prithvi R., Ekaterina K., Smestad, B., Asbjørnslett B., Bhattacharyya, A. 2021. Predicting vessel speed in the Arctic without knowing ice conditions using AIS data and decision trees.

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. “Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Russel, S & Norvig, P. 2016. Artificial Intelligence: A Modern Approach.

Rudin, C. (2018) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

Saaranen-Kauppinen, A & Puusniekka, A. 2006. KvaliMOTV - Menetelmäopetuksen tietovaranto. Tampere: Yhteiskuntatieteellinen tietoarkisto. <https://www.fsd.tuni.fi/menetelmaopetus/>. (Viitattu 23.6.2021)

Salminen, A. 2011. Mikä kirjallisuuskatsaus? Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin.

Sarker, I. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. Saatavilla: <https://doi.org/10.1007/s42979-021-00592-x>

Saunders, M., Lewis, P. & Thornhill, A. 2009. Research methods for business students. 5. painos.

Shalew-Shwartz, S & Ben-David, S. 2014. Understanding Machine Learning. Saatavilla: <https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf> (Viitattu 5.11.2021)

Siren, T & Pekkarinen, O. 2017. TIETEENFILOSOFIS-METODOLOGISIA PERUSTEITA PRO GRADU -TUTKIELMAN LAADINTAAN. ISSN 2489-2769

Siukonen, T. & Neittaanmäki, P. (2019) Mitä tulisi tietää tekoälystä. Jyväskylä: Docendo.

Sivadas, A & Samuel, A. 2019. Artificial Intelligence and the marine industry. Saatavilla: <https://www.wartsila.com/insights/article/artificial-intelligence-and-the-marine-industry> (Viitattu 29.12.2021)

Taulli, T. 2019. Artificial Intelligence Basics: A Non-Technical Introduction.

Toivola, J. 2016. Winternavigation and Icebreaking services, Baltic countries co-operation Urban Node 2016. <https://vayla.fi/documents/25230764/0/Jarkko+Toi-%20vola.pdf/a1944554-cfba-4127-ac2f-17fb4051907b> (Viitattu 20.6.2021)

Traficom. 2022. Alusten jääluokat. <https://www.traficom.fi/fi/liikenne/merenkulku/alusten-jaaluokat> (Viitattu 20.2.2022)

Turing, A. 1950. Computing Machinery and Intelligence.

Viljanen, M. 2021. Pitääkö EU:n AI-asetusehdotuksesta olla huolissaan? <https://etairos.fi/2021/05/17/pitaako-eun-ai-asetusehdotuksesta-olla-huolissaan> (Viitattu 5.8.2021)

Väylävirasto. 2020. Suomen talvimerenkulku. Saatavilla: <https://vayla.fi/documents/25230764/43984491/Suomen_talvimerenkulku_2020-2021_FI_4.11.2020.pdf/be936682-29a6-5353-8dbb-14e3dcfe5cff/Suomen_talvimerenkulku_2020-2021_FI_4.11.2020.pdf> (Viitattu: 1.9.2021)

Zhang, K. et al. 2020. Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control.