Koneoppimismenetelmien hyödyntäminen jääkiekkodatan avulla tehtäviin ottelu- ja pelaajaennusteisiin

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Joulukuu 2024
Santeri Isotalo

TURUN YLIOPISTO
Tietotekniikan laitos

Santeri Isotalo: Koneoppimismenetelmien hyödyntäminen jääkiekkodatan avulla tehtäviin ottelu- ja pelaajaennusteisiin

LuK-tutkielma, 34 s.
Tietojenkäsittelytiede
Joulukuu 2024

Jääkiekko on yksi suosituimmista urheilulajeista etenkin Pohjois-Amerikassa ja Pohjoismaissa. Jääkiekko on nopeavauhtinen peli, jossa tapahtuu useita erilaisia tapahtumia, kuten laukauksia, syöttöjä, taklauksia ja tappeluita. Lajin fyysisyys johtaa siihen, että useat pelaajat loukkaantuvat uransa aikana esimerkiksi aivotärähdykseen. Otteluissa satunnaisuudella on suuri rooli tapahtumien kulussa, ja vaikka jääkiekosta on tarjolla runsaasti erilaisia tilastoja, ei tilastoista suoraan voida päätellä mitä tulevissa otteluissa tulee tapahtumaan. Tämän takia tapahtumien ennustamiseen on pyritty keksimään kehittyneempiä menetelmiä, kuten koneoppimisalgoritmien hyödyntämistä. Koneoppiminen on yksi tekoälyn osa-alueista, jossa algoritmit pyrkivät ratkaisemaan ongelmia datan avulla ilman suoria ratkaisuohjeita, minkä takia koneoppimismalleja voi käyttää myös monimutkaisiin ongelmiin.
Selvitin, kuinka koneoppimismenetelmiä voi hyödyntää jääkiekkodatan avulla tehtäviin ennusteisiin, kuinka mallit suoriutuivat sekä mitä piirteitä malleissa hyödynnettiin. Suurin osa aiemmasta kirjallisuudesta onnistui ennustamaan jääkiekko-otteluiden voittajia hieman alle 60 % tarkkuudella. Osa aiemmista menetelmistä pääsi noin 77 % tarkkuuteen ja yksi tutkimus sai testidatalle jopa 91,8 % tarkkuuden. Pelaajien loukkaantumisten osalta aivotärähdysten ennusteissa päästiin AUC:lla mitattuna parhaimmillaan arvoon 0,79, mikä on parempi kuin NHL:n ottelutarkkailijoiden tarkkuus. Tulevien kausien loukkaantumisia ennustettiin tarkasti, noin 95 % tarkkuudella. Myös pelaajien ja maalivahtien luokitteluun saatiin tarkkoja tuloksia koneoppimismenetelmillä.
Suurin osa aiheen aiemmasta kirjallisuudesta koski maailman suosituinta jääkiekkoliigaa NHL:ää. Jatkossa tutkimuksia voisi kohdentaa pienempiin eri maiden liigoihin, jotta aiemman kirjallisuuden yleistyttävyys selkeytyisi. Lisäksi aiemman kirjallisuuden tutkimukset koostuivat perinteisemmistä koneoppimismalleista, joten tulevaisuudessa esimerkiksi neuroverkkojen hyödyntäminen voisi lisätä ennusteiden tarkkuutta.

Asiasanat: koneoppiminen, jääkiekko, algoritmit, suorituskykymittarit

Sisällys

Kuvat

Taulukot

Termistö

ANP Analytical Network Process

API Application Programming Interface, ohjelmointirajapinta

AUPRC Area Under the Precision-Recall Curve, alue tarkkuus-herkkyyskäyrän alla

AUROC Area Under the Receiver Operating Charasteric, alue ROC-käyrän alla

CNN Convolutional Neural Network, konvoluutioneuroverkko

CSS Central Scouting Service, kykyjenetsintäpalvelu

GIM Goal Impact Metric

KNN K-Nearest Neighbours, K-lähintä naapuria

LSTM Long Short-Term Memory

NHL National Hockey League

NLP Natural Language Processing, luonnollisen kielen käsittely

RNN Recurrent Neural Network, takaisinkytketty neuroverkko

ROC Receiver Operating Characteristic

TF-IDF Term Frequency/Inverse Documentation Frequency, termin tärkeys suhteessa sen esiintymistiheyteen dokumentaatiossa

1 Johdanto

Mediassa esiintyy toisinaan väite, että jääkiekkoa ei voida ennustaa. Esimerkiksi suomalaisen jääkiekkoliigan podcastissa analyytikko Matti Honkanen kertoo pitkäaikaisen valmentajan Risto Dufvan kysyneen "Mikä on kaikista suurin ratkaiseva tekijä, miten peli päättyy? Tuuri."[1]. Ei ole epätavallista, että tilastojen valossa heikompi joukkue voittaa lopulta ottelun. Esimerkiksi 13.11.2024 pelattu National Hockey Leaguen (NHL) ottelu Carolina Hurricanes vastaan Utah Hockey Club päättyi Utahin 1–4 voittoon, vaikka NHL-ennustamiseen erikoistuneen MoneyPuck-sivuston mukaan Carolina olisi ottelun jälkeisten tilastojen perusteella ansainnut voittaa ottelun 97,1 % todennäköisyydellä[2].

Sattuman suuri merkitys jääkiekko-otteluiden lopputuloksille on tunnettu jo pitkään. The Success Equation -kirja[3] tuo esiin tilastojen variansseja hyödyntämällä, että tuuri vaikuttaa NHL-otteluissa 53 %, mikä on selkeästi enemmän kuin esimerkiksi jalkapallon Valioliigan tuurin osuus 34 % tai NBA-koripalloliigan tuurin osuus 12 %. Kirjassa tuodaan esiin, että tuurin osuus on samankaltainen muissakin näiden lajien ammattilaisliigoissa. Tätä puoltaa tutkimus[4], jossa todettiin, että NHL:ssä 76 % otteluista ratkeaa sattuman kautta, jolloin otteluiden ennustusten tarkkuus voi olla korkeintaan 62 %. Myös toista näkökulmaa ennusteisiin löytyy, sillä kotimaisessa Liigassa suomalainen yhtiö Digia Oyj kertoo osuneensa tekoälyn avulla oikeaan jopa 76 % otteluista[5].

Tekoälyn suosio on viime vuosina kasvanut etenkin generatiivisten mallien an-

siosta, ja Menonin[6] mukaan nykyään "lähes kaikki käyttävät tekoälyä joko tietoisesti tai tiedostamatta". Koneoppiminen on yksi tekoälyn osa-alueista ja monia koneoppimisen algoritmejä käytetäänkin jo onnistuneesti ennustamaan esimerkiksi erilaisia sairauksia terveydenhuollossa[7]. Koneoppimisennustuksiin tarvitaan paljon dataa, jota jääkiekosta on tarjolla esimerkiksi eri jääkiekkoliigojen verkkosivuilla. Jos koneoppimismallille syötetään riittävästi oikeanlaista dataa, voitaisiin jääkiekosta mahdollisesti havaita sellaisia lainalaisuuksia, joita ei suoraan välttämättä itse havaittaisi ottelua seuraamalla. Perinteisempää tilastotiedettä on hyödynnetty esimerkiksi Poisson-jakauman avulla jääkiekko-otteluiden voittajia ennustaessa[8], joten myös kehittyneemmät koneoppimismenetelmät voisivat toimia ennustamiseen.

Riittävän suurella tarkkuudella osuvista koneoppimismalleista voitaisiin hyötyä sekä taloudellisesti että urheilullisesti. Pelaajien loukkaantumisen kesto voitaisiin arvioida esimerkiksi uutismedioissa ilman tarkkoja potilastietoja syöttämällä yleisesti tiedossa oleva informaatio koneoppimismallille. Joukkueen johdon ei tarvitsisi koneoppimismallien ansiosta nähdä jokaisen pelaajan jokaista ottelua voidakseen arvioida pelaajien sopivuutta joukkueeseen tai määrittääkseen pelaajien sopimusten rahallista arvoa. Tarkkaan osuvilla loukkaantumisennusteilla voitaisiin saada merkittävää taloudellista hyötyä ennaltaehkäisyn kautta. Esimerkiksi NHL:ssä on arvioitu loukkaantumisten maksavan vuodessa 218 miljoona dollaria, joista pelkät aivotärähdykset aiheuttavat jo 42,8 miljoonaa dollaria[9]. Taloudellista hyötyä olisi tarjolla myös, jos vedonlyöjä pystyy saavuttamaan mallin, joka laskee voittajien todennäköisyydet tarkasti. Vedonlyönti voi olla mahdollista positiivisella odotusarvolla[10], jos vedonlyöjällä on tiedossa tapahtumien todennäköisyydet.

Tutkielman tavoitteena on vastata seuraaviin tutkimuskysymyksiin:

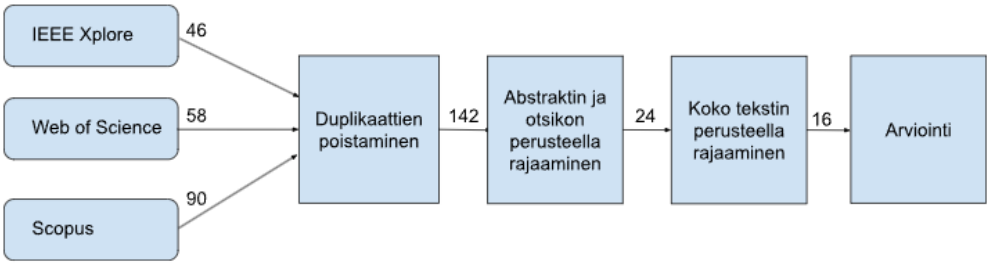
1. Mitä koneoppimismenetelmiä on tähän mennessä hyödynnetty jääkiekkodatan avulla tehtäviin ottelu- ja pelaajaennusteisiin?

2. Kuinka tarkasti koneoppimismallit pystyivät tekemään jääkiekkodatan avulla

ennusteita?

3. Mitkä tekijät vaikuttivat eniten mallin tekemään ennusteeseen?

Tutkielma toteutettiin kirjallisuuskatsauksena perustuen tieteellisiin lähteisiin. Käytetyt lähteet olivat ensisijaisesti IEEE Xplore-, Web of Science- ja Scopus-tietokannoista, joissa käytettiin hakulauseketta ("machine learning"OR "predictive model*") AND hockey. Haku tuotti 194 tulosta, jotka karsittiin kuvan 1.1 perusteella.



Kuva 1.1: Hakuprosessi

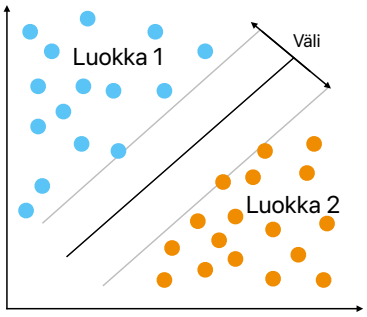
Poistamalla duplikaatit jäljelle jäi 142 hakutulosta. Otsikon ja abstraktin perusteella hakutuloksista pystyttiin rajaamaan 118 tulosta pois, sillä monet hakutuloksista käsittelivät esimerkiksi ilmakiekkoa, rullakiekkoa tai tappeluiden tunnistamista, jotka eivät ole oleellisia tämän tutkimuksen kannalta. Jäljelle jäi 24 lähdettä, jotka luettiin kokonaan. Koko tekstin perusteella jäi lopulta jäljelle 16 tekstiä. Lisäksi täydentäviä lähteitä on löydetty Google Scholarin ja Googlen hakukoneen avulla.

2 Koneoppimismenetelmät jääkiekossa

2.1 Koneoppimisen algoritmit

K-lähintä naapuria (K-Nearest Neighbours, KNN) on algoritmi, joka antaa datapisteelle luokan jo aiemmin luokiteltujen pisteiden läheisyyden perusteella[11]. Esimerkiksi 2-lähintä naapuria -algoritmissa data jaetaan eri luokkiin siten, että datapisteen kahden lähimmän olemassa olevan datapisteen luokan perusteella päätetään uudelle datapisteelle sopiva luokka. K-keskiarvot (K-means) -algoritmilla puolestaan tarkoitetaan luokittelualgoritmia, joka käyttää matemaattisia kaavoja minimoidakseen kahden datapisteen välisten keskipisteiden (centroid) etäisyyden[12]. Näiden etäisyyksien avulla datapisteet jaetaan luokkiin, jolloin lähellä toisiaan olevat datapisteet päätyvät samaan luokkaan. K-lähintä naapuria on siis ohjattua oppimista, sillä luokitteluun tarvitaan jo olemassa olevaa dataa, kun taas K-keskiarvot-algoritmi on ohjaamatonta oppimista, sillä luokittelu tapahtuu ilman olemassa olevaa luokittelua.

Tukivektorikone toimii käytännössä piirtäen marginaaleja datapisteiden väliin, jonka perusteella ennustettavat luokat muodostuvat (kuva 2.1) [13].



Kuva 2.1: Esimerkki tukivektorikoneen luokittelusta [13]

Tukivektorikoneen marginaali on jakanut datapisteet kahteen eri luokkaan, jotka ovat marginaalien ylä- ja alapuolella.

Päätöspuulla tarkoitetaan puun kaltaista luokittelijaa, jonka ylhäällä on juuri ja jossa edetään alaspäin kohti lehtiä. Puussa edetään alempiin lehtisolmuihin eri kriteerien perusteella, kunnes päädytään luokitteluun.[14] Satunnaismetsä koostuu useasta päätöspuusta, joiden ennusteet yhdistetään lopputulokseksi. XGBoost muodostaa useita puita, joissa kukin uusi puu korjaa aiempien puiden virheitä, muodostaen lopullisen ennusteen kaikkien muodostettujen puiden keskiarvon perusteella[15]. Nämä kaikki ovat ohjattua oppimista, sillä päätöspuu rakennetaan tiedossa olevan datan perusteella.

Naiivi Bayes-luokittelija hyödyntää tilastotieteestä tuttua Bayesin teoreemaa luokitellakseen datapisteen parhaalla mahdollisella tavalla[16]. Bayes-verkolla tarkoitetaan mallia, joka tunnistaa riippuvuuksia tapahtuneen tuloksen ja sen muuttujien välillä, pyrkien tunnistamaan mitkä muuttujat vaikuttivat lopputulokseen[17].

Logistinen regressiomalli laskee selitettävälle muuttujalle arvon 0 ja 1 välillä, jonka perusteella luokittelu tapahtuu[18]. Tämä pätee siis hyvin tapahtumiin, joissa on vain kaksi vaihtoehtoa, sillä arvo voidaan approksimoida ennusteeksi 0 tai 1. Tällä välillä saataville ennusteille saadaan lisäksi todennäköisyys, koska esimerkiksi arvo 0,65 voidaan tulkita 65 % todennäköisyytenä. Lineaarinen diskriminanttianalyysi

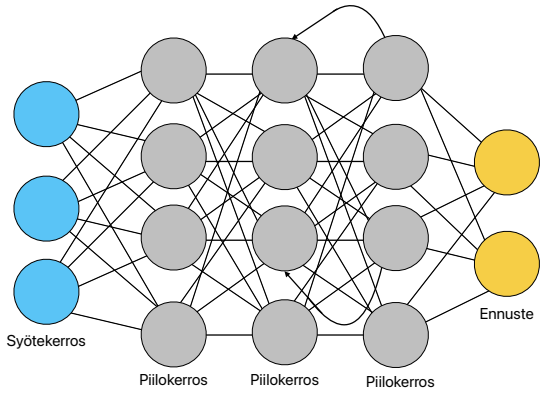
muodostaa piirteistä lineaarisia yhdistelmiä, joiden avulla havainnot erotetaan omiin luokkiinsa[18].

Neuroverkot ovat toisiinsa liitetty kokoelma matemaattisia malleja[19]. Neuroverkkoja on useita erilaisia ohjatun, puoliohjatun ja ohjaamattoman oppimisen menetelmiä, jotka soveltuvat käytettäväksi erilaisiin tilanteisiin[20]. Jääkiekkoennusteiden kannalta oleellisia ovat esimerkiksi konvoluutioneuroverkot (convolutional neural network, CNN), joita voidaan hyödyntää esimerkiksi kaukalokarttojen laukausten tunnistamiseen. Takaisinkytketty neuroverkko (recurrent neural network, RNN), "pitkä lyhyen aikavälin muistin neuroverkko"(long short-term memory, LSTM) sekä muuntimet (transformers) puolestaan voivat olla hyödyllisiä jääkiekosta tarjolla olevaan aikasarjadataan.

Konvoluutioneuroverkko koostuu useasta konvoluutiokerroksesta, jotka laskevat syötteen ja painokertoimien avulla tulon. Nämä tulot etenevät aktivaatiokerrokseen, jossa käytetään epälineaarisuutta konvoluutiokerrosten yksinkertaistamiseen. Yhdistämiskerros (pooling layer) pienentää dataa entisestään säilyttäen tärkeimmät piirteet. Lopulta täysin yhdistetty kerros (fully connected layer) muodostaa näistä piirteistä ennusteen.[20]

Takaisinkytketyssä neuroverkossa on mahdollista palata kerrosten välillä taaksepäin[21]. Algoritmin ennusteisiin vaikuttavat sen hetkisen kerroksen lisäksi myös aiempien kerrosten syötteet. Takaisinkytketyt neuroverkot kärsivät häviävän gradientin ongelmasta, jossa useiden toistojen jälkeen algoritmi päivittää painokertoimia siten, että verkko ei lopulta hyödynnä pidemmän aikavälin tietoja[22]. LSTM-verkko pyrkii ratkaisemaan tämän ongelman muistisoluilla (memory cell) sekä porttiyksiköillä (gate units). Porttiyksiköt hallitsevat tiedonkulkua ja painokertoimien säätämistä siten, että gradientit eivät häviä ja muistisolut säilyttävät tietoa pidemmän aikaa.[23] Muuntimet liikkuvat kerrosten välillä huomiomekanismilla ja algoritmi oppii syötteestä, mikä tieto on tärkeää suhteessa muuhun tietoon[24]. Muunnin voisi

esimerkiksi ennustaa ottelun maaleja oppimalla huomioimaan kaikki ne tapahtumat, jotka vaikuttavat maalien syntymiseen.



Kuva 2.2: Esimerkki takaisinkytketyistä neuroverkoista

Takaisinkytketyssä neuroverkossa (kuva 2.2) kerrokset yhdistyvät toisiinsa lineaarisella yhdistelmällä, mutta aktivaatiofunktio lisää niihin epälineaarisuutta[18]. Syötekerros syöttää verkostoon dataa, jota piilokerrokset muokkaavat aktivaatiofunktioilla, kunnes lopulta annetaan ennuste[20].

Bagging ja boosting ovat yleisesti tunnettuja yhdistelmäalgoritmeja, joiden on todettu olevan tehokkaita algoritmien yleistämisessä. Baggingin avulla muodostetaan useita malleja erilaisilla bootstrapping-menetelmän avulla kerätyillä otoksilla datasta ja lopulta näiden mallien tulokset yhdistetään ennusteen saamiseksi. Boostingissa puolestaan koulutetaan useita eri malleja, joissa malli oppii aina aikaisemmasta mallista, painottaen koulutuksessa etenkin aikaisemman mallin virheitä. Lopulta nämäkin mallit yhdistetään ennusteen saamiseksi.[25]

Koneoppimismallit voivat ylioppia (overfit), jolla tarkoitetaan sitä, että malli ei toimi yleistettävästi uuteen dataan. Ylioppinut malli voi toimia todella hyvin sille syötetyllä koulutusdatalla, mutta testidatan kanssa toimivuus onkin huono. Tämä johtuu siitä, että ylioppinut malli on oppinut muistamaan koulutusdataa ulkoa sekä malli on sovittunut datan satunnaiseen kohinaan (noise) sen sijaan, että se olisi oppinut yleistettävän periaatteen datan takana.[26] Datavuodolla tarkoitetaan sitä,

että malli saa dataa, jota sen ei ole tarkoitus saada[27]. Tämä johtaa siihen, että malli ennustaa liian hyvin, koska se tietää jotain, mitä sen ei pitäisi tietää. Jos esimerkiksi joukkuekohtaiset maalimäärät ovat saatavilla mallin muuttujissa, malli saa ennustettua voittavan joukkueen täydellisellä tarkkuudella.

2.2 Koneoppimismallien arviointi

Koneoppimismallien toimivuutta voidaan mitata usealla eri suorituskykymittarilla. Yksinkertaisin tapa mitata toimivuutta on tarkkuus, jolla tarkoitetaan sitä, kuinka usein malli osuu oikeaan[28]. Jos esimerkiksi testidatassa on 100 kohdetta ja malli ennustaa niistä 95 oikein, on mallin tarkkuus 95 %. Tämä ei kuitenkaan tarkoita, että malli on välttämättä käyttökelpoinen. Jos datassa on esimerkiksi 95 positiivista lukua ja malli ennustaa aina positiivista, saadaan korkea tarkkuus ilman, että mallia voidaan kuitenkaan pitää yleisesti käyttökelpoisena.

Ennustukset voivat olla joko oikeita positiivisia (true positive), oikeita negatiivisia (true negative), vääriä positiivia (false positive) tai vääriä negatiivisia (false negative). Oikea positiivinen tai negatiivinen ilmaisee sitä, että malli on ennustanut oikein. Väärä positiivinen tai negatiivinen ilmaisee sitä, että malli on ennustanut väärin. Näistä voidaan laskea yleisiä tarkkuusmittareita seuraavilla laskukaavoilla [28]:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

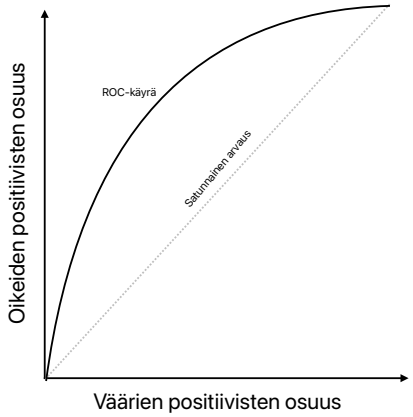
$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Herkkyys (recall) kuvaa sitä, kuinka usein malli tunnistaa oikeat positiiviset tapaukset. Tarkkuus (precision) kuvaa sitä, kuinka usein malli välttää ennustamasta väärää positiivista. Näiden lisäksi voitaisiin laskea myös spesifisyys (specificity), joka

on sama asia kuin herkkyys, mutta negatiivisille arvoille. Näitä kaavoja hyödyntämällä voidaan laskea myös F1-arvot alla olevalla kaavalla.[28]

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-arvon kaavasta havaitaan, että se antaa suuria lukuja silloin, kun malli suoriutuu hyvin sekä tarkkuudessa että herkkyydessä, koska näiden kertolasku on osoittajassa. Yleisiä mittareita ovat myös ROC-käyrä (Receiver Operating Characteristic) sekä alue ROC-käyrän alla (Area Under the Receiver Operating Charasteric, AUROC)[28].



Kuva 2.3: Esimerkki ROC-käyrästä

ROC-käyrä (kuva 2.3) kuvaa oikeiden positiivisten osuutta väärien positiivisten osuuteen. Kaikki käyrän alapuolelle jäävä on aluetta ROC-käyrän alla, eli AUROC. Mitä tarkempi malli on, sitä enemmän se ennustaa oikeita positiivia ja vähemmän vääriä positiivisia. Täydellisessä mallissa ROC-käyrä nousisi heti vasempaan yläkulmaan, jolloin sen alapuolelle jäävä alue vastaisi arvoa 1, eli näin ollen AUROC saa arvon välillä 0–1. Mitä suurempi arvo on, sitä paremmin ennustava malli on kyseessä.

Tämän lisäksi on olemassa myös tarkkuus-herkkyyskäyrän alle jäävä alue (Area Under the Precision-Recall Curve, AUPRC). Tämä eroaa AUROC-käyrästä siten, et-

tä AUROC painottaa kaikkia vääriä positiivisia tasaisesti, kun taas AUPRC painottaa enemmän tietyn kynnyksen ylittäviä vääriä positiivisia. Näin ollen AUPRC sopii tilanteisiin, joissa halutaan priorisoida mallin kykyä tunnistaa positiivinen luokka, kun taas AUROC sopii tilanteisiin, joissa kaikki virheet ovat yhtä tärkeitä.[29] Eri malleissa korostuvat eri suorituskykymittareiden tärkeys. Sähköpostin roskapostisuodattimessa on tärkeää saada suuri tarkkuus, eli välttää ei-roskapostien leimaaminen roskapostiksi, kun taas lääketieteessä tärkeämpää on saada korkea herkkyys, eli tunnistaa esimerkiksi sairaudet riittävän usein.

Häviöfunktioilla (loss functions) on tärkeä rooli koneoppimismalleja rakentaessa ja niiden suorituskykyä parannettaessa. Häviöfunktiot mittaavat kuinka suuri virhe ennustetulla arvolla ja todellisella arvolla on. Näitä käytetään etenkin regressiopohjaisissa koneoppimismalleissa. Yksi tunnetuimmista häviöfunktioista on neliövirhe, joka laskee todellisen ja ennustetun arvon välisen neliöjuuren.[30]

Koneoppimismalleja arvioidessa niiden käyttökelpoisuutta voidaan verrata toisiinsa muidenkin kuin suorituskykymittareiden perusteella. Esimerkiksi mallin läpinäkyvyys voi olla tärkeää, jos halutaan tietää mistä mallin ennuste muodostuu[31]. Myös mallien suorituskyky voi olla ratkaiseva tekijä – yksi päätöspuu on laskennallisesti yksinkertaisempi muodostaa, kuin tuhansien puiden satunnaismetsä.

2.3 Jääkiekko

Jääkiekko on liikevaihdoltaan yksi maailmanlaajuisesti suurimmista urheilulajeista ja esimerkiksi NHL:n liikevaihto oli kaudella 2023–2024 arviolta noin 6,2 miljardia dollaria[32]. Jääkiekossa joukkueiden tavoitteena on yksinkertaisesti tehdä enemmän maaleja kuin vastustajan joukkue. Jääkiekkoliiton sääntöjen mukaan[33] ottelussa pelataan kolme 20 minuutin erää siten, että molemmilta joukkueilta on kentällä samaan aikaan kuusi pelaajaa, yleensä viisi näistä on kenttäpelaajia ja yksi maalivahti. Se joukkue, jolla on enemmän maaleja kolmen erän jälkeen voittaa. Jos tilanne on

tasan, ottelu etenee viiden minuutin jatkoaikaan, jossa joukkueilla on kentällä vain kolme pelaajaa ja maalivahti. Jatkoajalla voittaja on se, joka saa kiekon ensimmäisenä maaliin. Jos kumpikaan joukkue ei voita ottelua kolmen erän tai jatkoajan aikana, siirrytään voittolaukauskilpailuun, jossa vuorotellen joukkueesta lähetetään yksi pelaaja yrittämään maalintekoa vain vastustajan maalivahdin vartioimaan maaliin. Voittolaukauskilpailussa selvitetään ensin, kumpi joukkue on parempi kolmen laukauskierroksen aikana, mutta jos tilanne on vielä tämänkin jälkeen tasan, voittaja ratkaistaan äkkikuolemalla. Ottelun voittaa äkkikuolemassa se joukkue, joka tekee maalin, jos vastustaja ei onnistu samalla kierroksella myös maalinteossa.

Jääkiekko on fyysinen laji ja sen säännöt[33] sallivat kovat taklaukset ja jopa nyrkkitappelut tietyin ehdoin. Jos taklaus tehdään väärin, esimerkiksi se osuu selkään tai päähän, saa taklaamisesta rangaistuksen. Jos rike on pieni, rangaistus on yleensä kaksi minuuttia, jolloin rangaistu joukkue saa laittaa maalivahdin lisäksi kentälle vain neljä pelaajaa viiden sijaan. Jos rikkeitä tehdään samaan aikaan useampi, voi minimissään joukkueella olla vain kolme pelaajaa kentällä maalivahdin lisäksi. Jos rike on suuri se voi johtaa siihen, että rikkeen tehnyt pelaaja saa ottelun loppuun asti pelikiellon ja mahdollisesti jopa tuleviin otteluihin kestävän kiellon.

Otteluiden fyysisyys johtaa siihen, että pelaajat voivat helposti saada fyysisiä vammoja. Tämän takia esimerkiksi NHL on palkannut erikseen tarkkailijoita (spotters), joiden tehtävä on seurata ottelun aikana tapahtuvia fyysisiä tapahtumia ja käskeä pelaaja pois ottelusta, jos hänellä epäillään mahdollista aivotärähdystä [31]. Otteluita pelataan kauden aikana runsaasti. Kausi kestää yleensä noin 8-9 kuukautta, jonka aikana esimerkiksi NHL:ssä pelataan 82 runkosarjaottelua per joukkue, jonka jälkeen hyvin menestyvät joukkueet pelaavat vielä pudotuspelejä, joilla selvitetään mikä joukkue on lopulta paras. Tämän lisäksi tulee vielä useiden tuhansien kilometrien matkustamiset sekä pelipäivien ulkopuolella tehtävät harjoitukset. Tämä kaikki aiheuttaa fyysistä rasitusta pelaajille.

3 Koneoppimismenetelmillä tehtävät ennustukset jääkiekodatasta

Jääkiekosta kerätään paljon dataa, jota on saatavilla eri liigojen verkkosivuilta. Mitä suurempi liiga on kyseessä, sitä enemmän tilastoja on tarjolla. Näitä tilastoja löytyy sekä joukkuekohtaisesti että pelaajakohtaisesti, joten yksi ottelu tuottaa jo useita kymmeniä eri tilastoja, minkä lisäksi tilastoja on myös kausikohtaisesti. Vaikka jääkiekkodataa on tarjolla todella runsaasti, ei ennustusten tekeminen silti ole helppoa. Tiuhaan tehtävät pelaajien vaihdot, korkeat nopeudet ottelun tapahtumissa, yhteentörmäykset ja tappelut tekevät jääkiekosta erityisen hankalan pelin ennustaa[34].

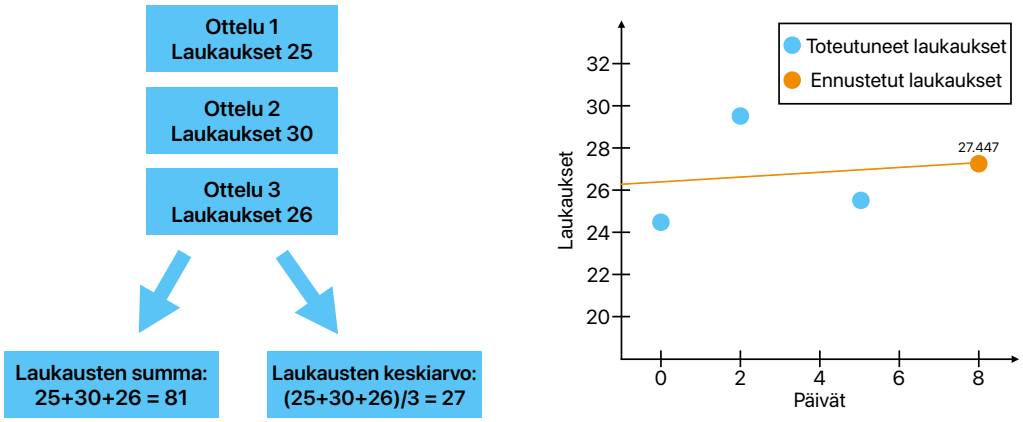
3.1 Ennusteet joukkueiden menestymiselle

Joukkueen menestymistä yksittäisissä otteluissa on pyritty tutkimaan monesta eri näkökulmasta. Jääkiekon ennustamiselle on ajateltu olevan tietty yläraja, jota paremmin otteluita ei voida ennustaa[4]. Artikkelissa[4] päädyttiin kausien 2005–2006 ja 2011–2012 välisten otteluiden voittojen keskihajontoja käyttämällä tulokseen, että NHL-otteluista 24 % olivat taidosta kiinni ja loput 76 % olivat tuurista kiinni. Tämän jälkeen saatiin vielä tilastollisella testillä paras todennäköisyys sille, että teoreettinen yläraja otteluiden ennustamisen tarkkuudelle on 62 %. Kyseinen tutkimus pyrki myös ennustamaan otteluita käyttäen apuna sekä numeerisia tilastoja

että ennen otteluita kirjoitettuja raportteja. Raporteista kerättiin sanoja käyttäen luonnollisen kielen käsittelystä (Natural Language Processing, NLP) tuttua menetelmää termin tärkeydestä suhteessa sen esiintymistiheyteen dokumentaatiossa (Term Frequency/Inverse Documentation Frequency, TF-IDF). Tämän lisäksi raportteihin käytettiin myös tunneanalyysiä (sentiment analysis), jonka avulla raporteista muodostettiin kolme eri muuttujaa, jotka olivat positiivisten sanojen lukumäärä, negatiivisten sanojen lukumäärä, sekä positiivisten ja negatiivisten sanojen erotuksen osuus kaikista sanoista. Malli koulutettiin 67 % koulutusdatalla ja sen toimivuutta kokeiltiin 33 % testidatalla. Usealla eri algoritmillä pyrittiin ennustamaan otteluiden tuloksia, käyttäen pelkästään tilastollista dataa, käyttämällä pelkästään TF-IDF:n avulla löydettyjä sanoja, käyttämällä pelkästään tunnepohjaisia sanoja, sekä käyttäen näiden kaikkien yhdistelmiä. Parhaiten pärjäsi yhdistelmä, jossa yhdistettiin numeeriset mallit, sanat sekä tunneanalyysi, jonka jälkeen mallien enemmistö valitsi voittajan oikein 60,25 % tarkkuudella. Kaikkien eri algoritmien tarkkuus oli välillä 50,20 % ja 60,25 %, joka puoltaisi artikkelin ajatusta siitä, että jääkiekko-otteluiden ennustamisen yläraja on lähellä 62 %. Artikkeli toi esiin myös sen, että kotijoukkueet voittavat otteluista 56 %, joten tätä huonommin pärjäävät algoritmit ovat huonompia ennustamiseen kuin se, että ennustaisi aina vain kotijoukkuetta.

Artikkeli[35] sai samankaltaisia tuloksia, mutta eri lähtökohdista. Artikkelissa ennustettiin otteluita käyttäen sekä momentumia että saatavilla olevia tilastoja. Momentum määriteltiin artikkelissa "joukkueen yleinen pelien laadun nouseminen tai laskeminen pienikokoisissa otteluotoksissa", eli katsotaan esimerkiksi viimeisen kolmen ottelun aikana tapahtuneita muutoksia. Data oli kerätty vuosien 2011 ja 2020 väliltä NHL:n julkisesta ohjelmointirajapinnasta (Application Programming Interface, API), ja otteluista oli valittu vain ne, jotka loppuivat ennen jatkoaikaa. Tämän lisäksi datasta rajattiin pois jokaisen joukkueen kauden ensimmäiset 20 ottelua, koska artikkelin mukaan otteluissa ei ole johdonmukaisuutta kauden alussa. Datas-

sa käytettiin n-otteluiden intervallia, jossa n tarkoittaa otteluiden lukumäärää. Koneoppimismallien piirteet olivat kolmenlaisia: frekvenssipohjaisia (frequency-based), momentumpohjaisia ja näiden kombinaatioita. Frekvenssipohjaisilla tarkoitettiin eri numeeristen tilastojen, kuten laukausten lukumäärien keskiarvoja tai yhteenlasketuja summia edellisestä n-ottelusta. Momentumpohjaiset tilastot olivat myös numeerisia, mutta tilastot olivat laitettu siten, että ne huomioivat myös päivien kulumisen, eikä vain edellisten n-otteluiden määrää (kuva 3.1). Historiallisen datan lisäksi momentumpohjaiset piirteet pyrkivät myös ennustamaan seuraavan tulevan ottelun tilastoja trendiviivalla.



(a) Frekvenssipohjainen data (b) Momentumpohjainen data

Kuva 3.1: Esimerkki frekvenssi- ja momentum-pohjaisesta datasta [35]

Artikkelissa data oli jaettu 80 % koulutusdataan ja 20 % testidataan. Muuttujat muutettiin siten, että sen sijaan, että olisi tehty sekä koti- että vierasjoukkueelle omat muuttujat, tehtiin muuttujat koti- ja vierasjoukkueen muuttujien erotuksesta. Koneoppimismalleja artikkelissa koulutettiin kolme erilaista: logistinen regressio, satunnaismetsä sekä lineaarinen diskriminanttianalyysi. Artikkelissa mainitaan myös, että XGBoostia ja neuroverkkoja yritettiin, mutta niillä saatiin vain harhaisia (biased) tuloksia, joten ne jätettiin pois tarkasteluista. Eri koneoppimismallit suoriutuivat eri tavalla eri n-intervalleilla (taulukko 3.1).

Taulukko 3.1: Eri ottelumäärien parhaimmat koneoppimismallit [35]

Otteluiden määrä	Koneoppimismalli	Parhaan mallin piirteet	Tarkkuus
3	LDA	Momentum	0,5661
4	Logistinen regressio	Piirteiden yhdistelmä	0,5608
5	Logistinen regressio	Momentum	0,5682
6	LDA	Frekvenssi	0,5691
7	LDA	Frekvenssi	0,5753
8	LDA	Frekvenssi	0,5829
9	Logistinen regressio	Frekvenssi	0,5889
10	LDA	Frekvenssi	0,5931

Momentumin hyödyntäminen piirteissä vaikutti silloin, kun otteluita valittiin vain vähän, kolmen ja viiden väliltä. Tämän jälkeen parhaissa malleissa korostuivat frekvenssipiirteet. Kolmesta koneoppimismallista parhaiten suoriutuivat sekä lineaarinen diskriminanttianalyysi että logistinen regressiomalli, mutta tarkempia tuloksia artikkelista katsoessa jokainen malli antaa tarkkuuksia 54,14 % ja 59,31 % väliltä, eli myös satunnaismetsä oli lähes yhtä toimiva vaihtoehto.

Konferenssijulkaisussa[36] otteluiden voittajia pyrittiin ennustamaan käyttämällä sekä tarjolla olevia perinteisiä tilastoja että "kehittyneempiä tilastoja"(advanced stats), kuten bloggaajien ja tilastotietelijöiden esiin tuomia tilastoja. Data oli kerätty kolmen kuukauden ajalta NHL-kauden 2012–2013 aikana ja sisälsi 517 ottelua. Osa tilastoista oli kerätty ennen ottelua, osa ottelun jälkeen. Ottelun jälkeen kerätyistä tilastoista saatiin muodostettua aikasarjadataa, koska datasta oli voitu laskea esimerkiksi edellisen kolmen ottelun keskiarvotilastoja, eikä vain tarjolla olleita koko kauden tilastoja. Artikkelissa mainitaan, että kehittyneitä tilastoja, kuten PDO (joukkueen laukaisuprosentin ja torjuntaprosentin summa) on vaikea kerätä, mutta julkaisu on vuodelta 2013 ja vuonna 2024 nämäkin tilastot ovat helposti internetistä

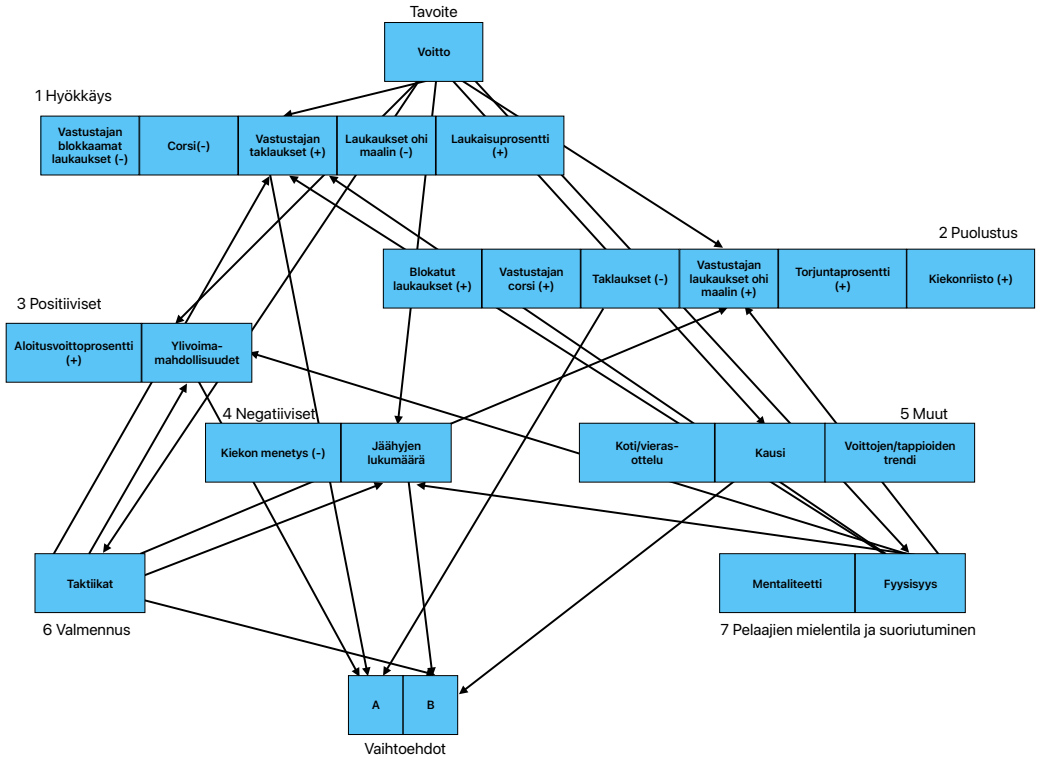
saatavilla. Julkaisu pyrki ennustamaan voittajan binääriluokittelulla, eli antamalla joukkuueelle arvoksi joko voiton tai häviön. Viittä eri koneoppimismallia koulutettiin 67 % koulutusdatalla ja 33 % testidatalla, käyttäen sekä perinteisiä tilastoja, kehittyneitä tilastoja, aikasarjadataa ja näiden yhdistelmiä. Parhaat tulokset saivat viiden piilokerroksen neuroverkot, tukivektorikone, sekä naiivi Bayes -luokittelija (taulukko 3.2).

Taulukko 3.2: Julkaisun [36] kolmen parhaan mallin suorituskykymittarit

Malli	Tarkkuus	Precision	Recall	F-arvo	ROC-käyrä
Neuroverkot	0,5938	0,583	0,594	0,588	0,600
Tukivektorikone	0,5861	0,586	0,586	0,586	0,586
Naiivi Bayes	0,5725	0,567	0,571	0,569	0,588

Neuroverkot olivat tarkin malli, mutta suorituskykymittareista nähdään, että malleilla ei ole suurta eroa. Julkaisussa[36] mainitaankin, että tukivektorikoneen ja neuroverkkojen välillä ei ole tilastollisesti merkittävää eroa. Tarkkuusluvut ovat samankaltaiset kuin kahdessa aiemmin mainitussa tutkimuksessa, ja ne pysyvät alle aiemmin arvioidun teoreettisen 62 % ylärajan.

Teoreettinen yläraja on kuitenkin pystytty ylittämään, esimerkiksi artikkeli[37] sai ennustuksille tarkkuuden 77,5 %. Artikkelissa kerättiin dataa 1230 ottelun tuloksista, tilastoista ja otteluiden yksityiskohdista, minkä lisäksi ennustamiseen käytettiin myös asiantuntijoiden arvioita. Tärkeimmät avainmuuttujat löydettiin tukivektorikoneen avulla, minkä lisäksi neljältä eri asiantuntijalta kysyttiin mitä mieltä he ovat vastakkain pelaavien joukkueiden eroista näiden muuttujien suhteen, esimerkiksi siis onko joukkueella A vai B parempi fyysisyys. Näistä tärkeimmistä muuttujista muodostettiin ANP-verkosto (Analytic Network Process)(kuva 3.2). ANP-verkosto sopii tilanteisiin, joissa päätöksiä ei voida tehdä hierarkkisesti, koska ylempien ja alempien tasojen muuttujat ovat toisistaan riippuvaisia[38].



Kuva 3.2: ANP-verkosto otteluiden ennustamiselle[37]

Ennustusten tekemiseen käytettiin seitsemää eri klusteria, jotka olivat hyökkäys, puolustus, positiiviset tekijät, negatiiviset tekijät, muut tekijät, valmennus ja pelaajien henkinen ja fyysinen suorituskyky. Nuolet kertovat, miten muuttujat ovat yhteyksissä toisiinsa. Vaihtoehdot A ja B olivat ottelussa pelaavat joukkueet, joille verkosto ennusti arvot nollan ja yhden välillä, suuremman arvon saanut joukkue ennustettiin voittajaksi[37]. Artikkelissa tätä ennustamisverkostoa käytettiin runkosarjan jälkeen pelatulle 89 pudotuspeliottelulle, jolloin osumatarkkuudeksi saatiin 77,5 %.

Konferenssijulkaisussa[39] pyrittiin ennustamaan joukkueiden tekemiä maalimääriä sekä ottelun voittajaa tai tasapeliä. Koneoppimismalleina käytettiin logistista regressiomallia, tukivektorikonetta sekä K-lähintä naapuria. Datana oli 33 eri NHL-joukkueen dataa kausien 2015-2021 väliltä. Mallien tulokset oli julkaistu konfuusiomatriiseina, joiden perusteella suorituskykymittarit on laskettu (taulukko 3.3).

Taulukko 3.3: Julkaisun[39] mallien suorituskykymittarit

Malli	Tarkkuus	Precision	Recall	F-arvo
Logistinen regressio	0,778	0,770	0,780	0,775
Tukivektorikone	0,777	0,778	0,768	0,773
K-lähintä naapuria	0,697	0,700	0,679	0,689

Logistinen regressio ja tukivektorikone saivat samankaltaisia suorituskykylukuja, kun taas K-lähintä naapuria ennusti otteluita heikommin.

Kaikkein parhaimman tarkkuuden sai artikkeli[34], jossa oli kerätty kausien 2007–08 ja 2016–17 välistä dataa. Artikkelissa käytettiin K-lähintä naapuria, tukivektorikonetta, naiivi Bayes -luokittelijaa, diskriminanttianalyysiä ja päätöspuita ennustamaan otteluiden voittajia. Artikkelissa tukivektorikone koulutettiin 1230 runkosarjaottelulla ja sen muuttujien kelpoisuutta testattiin selvittämällä, kuinka tarkka ennustus on käytettäessä mallia koulutusdatalle. Koulutusdatalle saatiin jopa 94,05 % tarkkuus, jonka jälkeen piirteitä lähdettiin testaamaan testidatalle eri koneoppismalleilla (taulukko 3.4).

Taulukko 3.4: Eri koneoppimismallien tarkkuuslukemat [34]

Luokittelija	Yhdistelmä	Koulutusdatan tarkkuus	Testidatan tarkkuus
K-lähintä naapuria	Ei	1,00	0,804
Tukivektorikone	Ei	0,952	0,906
Bayes	Ei	0,946	0,904
Diskriminantti	Ei	0,950	0,915
Diskriminantti	AdaBoost	0,946	0,915
Puu	AdaBoost	0,967	0,904
Diskriminantti	Bagging	0,950	0,911
Puu	Bagging	1,00	0,901
Puu	Boosting	0,988	0,914
Diskriminantti	RobustBoost	0,955	0,918
Puu	RobustBoost	0,969	0,904

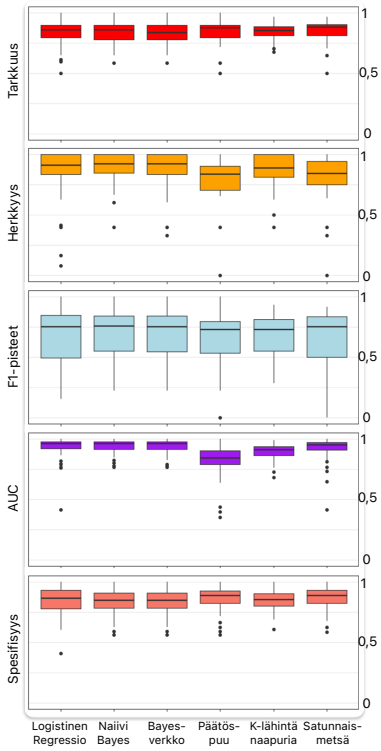
Kaikki luokittelijat antoivat todella korkeita tarkkuuksia, jotkut saivat koulutusdatassa jopa 100 % tarkkuuden. Testidatassa huonoin tarkkuus oli 80 %, mikä sekin oli parempi kuin yhdessäkään muussa tutkimuksessa. Tutkimuksessa avattiin tarkasti miten se toteutettiin, mutta näin korkeasta tarkkuudesta voisi epäillä kyseessä olleen esimerkiksi ylioppimista tai datavuotoa. K-lähintä naapuria tapauksessa voisi olla kyse ylioppimisesta, sillä koulutusdatassa tarkkuus oli jopa 100 %, kun testidatassa se oli vain 80,4 %. Muissa malleissa koulutusdata ja testidata olivat kuitenkin melko lähellä toisiaan, joten kyseessä voisi olla ollut datavuoto. Yksi mahdollinen virhe olisi voinut olla se, että jokaisessa ottelussa on aina kaksi joukkuetta, jolloin dataan saattaa olla jäänyt jokainen ottelu kaksi kertaa eri joukkueen näkökulmasta, jolloin malli tiesi vastauksen, jos se osasi yhdistää nämä tiedot toisiinsa. Vaikka mallit olivat tälle datalle todella tarkkoja, niiden yleistettävyys tuleville otteluille voi olla epäluotettavaa, jos tarkkuus johtui ei-yleistettävistä tekijöistä.

3.2 Pelaajien suoritusten ennustaminen

Pelaajien suoritusten ennustaminen on joukkueille tärkeää, jotta valmennusjohto osaa esimerkiksi valita toisilleen sopivat pelaajat jäälle ja laatia oikeansuuntaisia sopimuksia. Konferenssiartikkelissa[40] kerättiin NHL:n kaudelta 2015–2016 jopa yli kolme miljoonaa pelitapahtumaa, joiden perusteella koulutettiin viiden kerroksen neuroverkko, joka hyödynsi syväoppimista. Neuroverkko laski erilaisia Q-arvoja, joista muodostettiin lopulta pelaajille "maalivaikuttamismittari"(Goal Impact Metric, GIM), jonka perusteella pelaajat voitiin järjestää paremmuusjärjestykseen. Artikkelissa todettiin GIM-mittarin olevan toimiva, koska sen muodostaman järjestyksen kärjessä oli NHL:n tähtipelaajia, ja esimerkiksi sijalla kolme olevalla Johnny Gaudreaulla ja sijalla seitsemän olevalla Mark Scheifelellä oli alle miljoonan dollarin sopimus, mutta vuosi tämän tutkimuksen jälkeen molempien sopimus nousi yli viiteen miljoonaan dollariin. Vertauksessa GIM-mittari oli useaa muuta mittaria

parempi, ja sen korrelaatio seuraavan kauden pelaajasopimusten kanssa oli parhaimmillaan jopa 76,3 %.

Konferenssijulkaisussa[41] pelaajien luokitteluun käytettiin kuutta koneoppimismallia, jotka luokittelivat pelaajat kolmeen eri tasoluokkaan. Datana käytettiin kausien 2015–16 ja 2018–19 välistä dataa NHL:stä, minkä lisäksi pelaajien pisteytykseen otettiin Electronic Arts NHL -videopeleistä pisteitä samojen kausien ajalta. Pelaaja voi saada pisteitä asteikolla 1–99, joiden perusteella luokittelu tehdään parhaaseen 10 %, 25 % ja 50 % pelaajista. Käytetyt koneoppimismallit olivat logistinen regressio, K-lähintä naapuria, päätöspuu, satunnaismetsä, naiivi Bayes -luokittelija ja Bayes-verkko (kuva 3.3).



Kuva 3.3: Koneoppimismallien suorituskyky pelaajien luokittelulle [41]

Suurimmalle osalle datasta pelaajien luokittelussa parhaiten pärjäsi sekä naiivi Bayes -luokittelija että Bayes-verkko. Päätöspuilla ja satunnaismetsillä oli korkein

spesifisyys, mutta matala herkkyys. Parhaiten mallit ennustivat hyökkääjien luokittelua, mikä johtunee siitä, että tarjolla oleva data on usein hyökkäyspainotteista.[41]

Julkaisussa[42] käytettiin k-keskiarvot-algoritmia maalivahtien klusterointiin. Maalivahdit jaettiin viiteen eri ryhmään, jotka olivat riittämätön (inadequate), käyttökelpoinen (serviceable), kilpailukykyinen (competitive), eliitti (elite) ja maailmanluokka (world class). Suurin osa maalivahdeista olivat joko riittämättömiä tai maailmanluokkaa (taulukko 3.5).

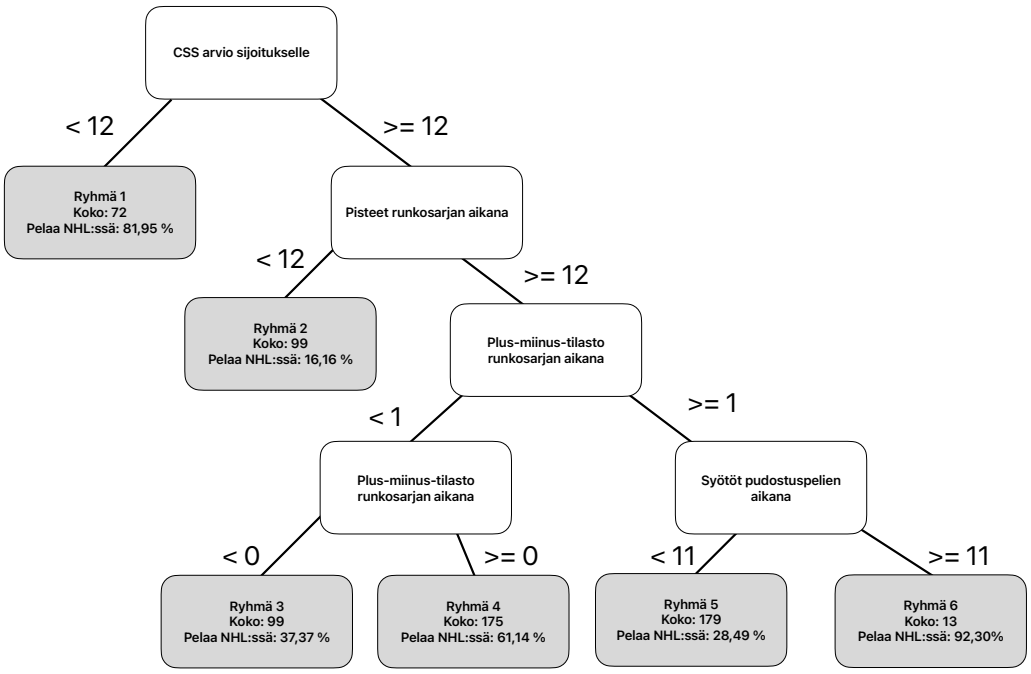
Taulukko 3.5: Maalivahtien luokittelu k-keskiarvot-algoritmilla[42]

Klusteri	Klusterin nimi	Määrä	Viiden vuoden kehitys
0	Riittämätön	10	3,2,2,1,1
1	Käyttökelpoinen	4	1,4,2,2,1
2	Kilpailukykyinen	4	4,4,1,2,1
3	Eliitti	8	3,4,4,2,1
4	Maailmanluokka	11	3,2,3,3,3

Viiden viime vuoden kehityksestä nähdään, että kaikki muut paitsi maailmanluokan maalivahdit ailahtelivat suorituksessaan melko paljon. Eliitti ja kilpailukykyiset maalivahdit aloittivat korkealla tasolla, mutta vuosien kuluessa laantuivat heikommaksi. Syy sille, miksi suurin osa maalivahdeista on joko todella hyviä tai todella huonoja algoritmin mielestä voi olla se, että jääkiekkojoukkueilla on ottelussa mukana kaksi maalivahtia, joista yleensä toinen on selkeä "ykkösmaalivahti". Esimerkiksi NHL-maalivahdeista kaudella 2023–2024 "ykkösmaalivahti"pelasi jopa 17 joukkueen kohdalla yli 60 % otteluista[43]. Ykkösmaalivahdin sopimus on usein myös kakkosmaalivahtia kalliimpi, esimerkiksi 71 % otteluista pelannut Juuse Saros tienasi kauden aikana viisi miljoonaa dollaria, kun joukkueen toinen maalivahti Kevin Lankinen tienasi 850 000 dollaria[44]. Näin ollen ykkösmaalivahti on oletettavasti kakkosmaalivahtia parempi, mikä jakaa jo maalivahdit selkeästi kahteen eri luokkaan.

NHL:ssä on joka vuosi varaustilaisuus, jossa joukkueet varaavat itselleen oikeuden pelaajiin. Usein nämä pelaajat ovat hetki sitten täysi-ikäistyneitä, jotka pelaa-

vat ympäri maailmaa erilaisissa matalamman tason liigoissa. Artikkelissa[45] pyrittiin identifioimaan poikkeukselliset pelaajat päätöspuiden avulla. Artikkelin data oli peräisin vuosien 1998-2008 varaustilaisuuksista, josta poistettiin maalivahdit. Data sisälsi pelaajista tietoa, kuten ikä ja pituus, sekä pelaajien suoriutumista otteluissa, kuten tehdyt maalit, pelatut ottelut ja peliaika. Tämän lisäksi dataa oli myös NHL:n kykyjenetsintäpalvelusta (Central Scouting Service, CSS), joka antaa pelaajille arvioita siitä, millä sijalla heidät tullaan varaamaan varaustilaisuudesssa. Artikkelissa tavoite oli ennustaa päätöspuilla tuleeko pelaaja pelaamaan yhtäkään ottelua NHL:ssä (kuva 3.4).



Kuva 3.4: Päätöspuu varaustilaisuuden pelaajien todennäköisyydelle pelata NHL:ssä [45]

Pelaajat, joiden CSS:n antama arvio oli alle 12, pelasivat 81,95 % todennäköisyydellä NHL:ssä. Tämän arvion lisäksi vaikuttavia tekijöitä olivat pisteet runkosarjan aikana, plus-miinus-tilasto runkosarjan aikana sekä syötöt pudotuspeleissä. Plus-miinus-tilastolla tarkoitetaan tilastoa, jossa pelaajan ollessa kentällä hän saa

pluspisteitä oman joukkueen tekemistä maaleista, kun taas vastustajan maaleista tulee miinuspisteitä. Pelaajat joilla oli positiivinen plus-miinus-tilasto ja yli 10 CSS-arvio pelasivat NHL:ssä jopa 92,30 % todennäköisyydellä. Artikkelissa tuotiin esiin neljän eri vuoden varaustilaisuus, jolloin mallin tarkkuus oli parhaimmillaan jopa 85,79%.

Koneoppimismalleja voidaan hyödyntää myös sopivien ketjujen muodostamiseen pelaajista. Konferenssiartikkelissa[46] tutkittiin, miten K-keskiarvot, satunnaismetsä ja tukivektorikone ennustavat kolmannen pelaajan lisäämisen vaikutusta kahden pelaajan rinnalle. Tutkimuksen kohteena oli kolmannen pelaajan vaikutus ketjun Corsi-prosenttiin. Corsi-prosentilla tarkoitetaan tilastoa, jossa verrataan tasalukuisin ketjuin tehtyjen oman joukkueen laukausyritysten määrän suhdetta sekä oman joukkueen että vastustajan joukkueen laukausyritysten määrään[47]. Artikkelissa[46] satunnaismetsää käyttämällä havaittiin, että kaikki muut tekijät vaikuttivat ketjun Corsi-prosenttiin enemmän kuin se, kuka kolmas pelaaja ketjuun lisätään. Pelaajat jaettiin K-keskiarvot-algoritmilla eri klustereihin, ja huomattiin että kolmas pelaaja aiheutti 50,37 % muutoksen ketjussa. Tukivektorikoneella saatiin 95 % tarkkuudella samat klusterit kuin k-keskiarvot-algoritmia käyttämällä. Artikkelin mukaan eniten klusterointiin vaikutti se, kuinka paljon kolmas pelaaja on aikaisemmin pelannut kahden muun pelaajan kanssa ja kuinka hyvin nämä kolme pelaajaa kontrolloivat irtokiekkoja.

3.3 Pelaajien terveyden ennustaminen

Jääkiekon fyysisyyden takia NHL on palkannut koulutettuja tarkkailijoita katsomaan jokaisen ottelun ja kirjaamaan ylös erilaiset tapahtumat, jotka saattaisivat viitata aivotärähdykseen[48]. Tälläisia tapahtumia ovat esimerkiksi jäässä liikkumatta makaaminen, tasapaino-ongelmat, tyhjä katse ja hitaasti taklauksen jälkeen jäältä nouseminen. NHL:n aivotärähdyksistä 46 % todetaan tarkkailijoiden katso-

mista tapahtumista, 54 % aivotärähdyksistä ei sisällä selkeitä nähtäviä merkkejä[31].

Artikkelissa[31] tutkittiin koneoppimismenetelmien hyödyntämistä aivotärähdysten ennustamisessa päätöspuun, satunnaismetsän sekä logistisen regression avulla. Data oli lähtöisin kausien 2018–2021 aikana tarkkailijoiden keräämästä tietokannasta. Lisäksi mahdollisia riskitekijöitä, kuten ikää ja aivotärähdyshistoriaa kerättiin NHL:n SCAT5®-tietokannasta. Tutkimuksessa saavutettiin päätöspuulla keskimäärin 0,76 AUC, kun taas logistisella regressiolla ja satunnaismetsällä saavutettiin keskimäärin 0,79 AUC. Logistisen regression ja satunnaismetsän kerrottiin toimivan tilastollisesti merkittävällä tavalla paremmin kuin päätöspuun, kun merkittävyyden rajana oli pidetty p-arvoa 0,05.

Taulukko 3.6: Aivotärähdysten ennustamisen tärkeimmät tekijät [31]

Logistinen regressio	Kerroin	Satunnaismetsä	Tärkeys
Liikkumattomuus	2,47	Tasapaino-ongelmat	100,00
Päähän kohdistunut lyönti	2,35	Liikkumattomuus	45,04
Tasapaino-ongelmat	2,09	Pelaaja pitelee päätänsä	5,73
Olkapää osunut päähän	1,14	Pää osunut jäähän	4,87
Pää osunut jäähän	0,73	Olkapää osunut päähän	4,18
Aiemmat aivotärähdykset	0,28	Hidas nouseminen	3,80
Erä	-0,39	Pää osunut laitaan	2,87
Pää osunut laitaan	-0,85	Erä	1,58
		Ikä	0,49
		Tyhjä katse	0,46
		Aiemmat aivotärähdykset	0,07
		Pää osunut pleksiin	0,00

Eri mallit painottivat eri tekijöitä (taulukko 3.6), esimerkiksi logistinen regressiomalli piti liikkumattomuutta tärkeimpänä tekijänä, kun taas satunnaismetsä piti tasapaino-ongelmia merkittävästi tärkeimpänä tekijänä. Logistinen regressiomalli piti päähän kohdistunutta lyöntiä toiseksi merkittävämpänä tekijänä, kun taas satunnaismetsä ei huomioinut sitä ollenkaan päätöksessään. Jotkut tekijät myös vaikuttivat ennusteeseen päinvastoin, esimerkiksi pään osuminen laitaan vähensi logistisen regressiomallin mielestä aivotärähdyksen todennäköisyyttä, satunnaismetsän mielestä se lisäsi aivotärähdyksen todennäköisyyttä. Artikkelissa tuotiin esiin, et-

tä logistinen regressiomalli antoi tarkkuudeltaan yhtä hyvän mallin kuin satunnaismetsät, mutta logistinen regressiomalli toi selkeämmin esiin mistä lopullinen ennuste muodostuu. Tämän selkeyden takia artikkeli suositteli logistisen regression valintaa tarkkailijoille. Lisäksi artikkelissa tuotiin esiin, että vaikka malli ennusti tarkkailijoita paremmin, ei sitä kannata käyttää ilman tarkempaa kontekstia pelaajasta, mutta se voi esimerkiksi parantaa tarkkailijoiden työtä poistamalla turhia pelinaikana tehtäviä terveystarkastuksia.

Myös artikkeli[49] toi esiin koneoppimismallien toimivuuden loukkaantumisennusteissa. Artikkelissa analysoitiin 2322 pelaajaa vuosilta 2007–2017 käyttäen logistista regressiota, satunnaismetsää, K-lähintä naapuria, naiivi Bayes -luokittelijaa, XGBoostia sekä kolmen suosituimman mallin yhdistelmämallia. Pelaajista 1317 loukkaantui tarkasteltujen vuosien aikana ja loukkaantumisia oli yhteensä 6673 kappaletta. Ennustettaessa seuraavan kauden loukkaantumisia artikkeli käytti jokaisen pelaajan viimeisimmän kauden loukkaantumis- ja suoriutumisdataa.

Taulukko 3.7: Koneoppimismallien tarkkuuslukuja seuraavan kauden loukkaantumisten ennustamiseen [49]

Malli	Tarkkuus	AUC	F1 pisteet
Kenttäpelaajat			
Logistinen regressio	$0,946 \pm 0,005$	$0,937 \pm 0,011$	$0,898 \pm 0,016$
Satunnaismetsä	$0,946 \pm 0,005$	$0,936 \pm 0,012$	$0,898 \pm 0,016$
K-lähintä naapuria	$0,700 \pm 0,020$	$0,752 \pm 0,028$	$0,577 \pm 0,036$
Naiivi Bayes	$0,854 \pm 0,027$	$0,917 \pm 0,015$	$0,775 \pm 0,035$
XGBoost	$0,946 \pm 0,005$	$0,948 \pm 0,010$	$0,898 \pm 0,016$
Top 3 yhdistelmä	$0,946 \pm 0,005$	$0,946 \pm 0,010$	$0,898 \pm 0,016$
Maalivahdit			
Logistinen regressio	$0,968 \pm 0,015$	$0,947 \pm 0,045$	$0,920 \pm 0,045$
Satunnaismetsä	$0,967 \pm 0,013$	$0,937 \pm 0,033$	$0,917 \pm 0,040$
K-lähintä naapuria	$0,808 \pm 0,041$	$0,816 \pm 0,076$	$0,618 \pm 0,105$
Naiivi Bayes	$0,943 \pm 0,023$	$0,936 \pm 0,031$	$0,869 \pm 0,054$
XGBoost	$0,967 \pm 0,013$	$0,956 \pm 0,026$	$0,917 \pm 0,040$
Top 3 yhdistelmä	$0,968 \pm 0,015$	$0,952 \pm 0,029$	$0,920 \pm 0,045$

Ennusteet (taulukko 3.7) osuivat paremmin maalivahtien kohdalla kuin kenttäpelaajien kohdalla. Usea malli ennusti melko samalla tavalla, mutta esimerkiksi K-

lähintä naapuria erosi selkeästi huonompana mallina muista, koska sen arvot olivat jokaikisessä sarakkeessa muita malleja huonompia. Paras malli kenttäpelaajille oli XGBoost, joka antoi tarkkuudeksi 94,6 %, minkä lisäksi sillä oli muita malleja parempi AUC. Maalivahtien kohdalla XGBoost-mallin tarkkuus oli 0,001 huonompi kuin logistisella regressiolla ja kolmen parhaan mallin yhdistelmällä, mutta sen AUC oli muita malleja parempi, joten XGBoost soveltui myös maalivahtien loukkaantumisen ennustamiseen parhaiten. Maalivahtien ja pelaajien ennusteissa korostuivat samat tekijät, sillä tärkeimpänä ennustustulokseen vaikuttivat aiempien loukkaantumisten määrä sekä ikä. Vaikka korkeat tarkkuuslukemat tarkoittavat, että koneoppimismallit ennustivat todella hyvin tulevia loukkaantumisia, voi tässäkin 94,6 % tarkkuus johtua ylioppimisesta tai datavuodosta. NHL-pelaajat myös loukkaantuvat todella usein, esimerkiksi keskushyökkääjistä 76,55 % loukkaantui vähintään kerran kausien 2007–2008 ja 2018–2019 välillä[50]. Voi siis olla, että koneoppimismalli loukkaantumisille ei ole helposti yleistettävissä tuleville vuosille, vaikka se onkin artikkelin[49] testidatalle ollut toimiva.

Pelaajien työmäärän aiheuttamasta rasituksesta ja palautumisesta on tehty tutkimus artikkelissa[51]. Artikkelissa seurattiin 17 tanskalaista ammattilaisurheilijaa jääkiekossa tarkkailemalla heidän sydämensykettä neljän viikon aikana, jolloin pelaajat harjoittelivat 23 kertaa ja pelasivat kahdeksan ottelua. Tämän jälkeen datapisteet klusteroitiin käyttäen K-keskiarvot++-algoritmia. Artikkelissa kerrottiin, että K-keskiarvot++-algoritmi eroaa perinteisestä K-keskiarvot-algoritmista siten, että satunnaisuuden sijaan se valitsee kaukaisimman mahdollisen etäisyyden perusteella ensimmäisen klusterin keskipisteen. Tutkimuksessa käytettiin erilaisia kiihdyttämisen, hidastamisen ja sydämen sykkeen avulla saatavia muuttujia, joiden perusteella K-keskiarvot++-algoritmi jakoi pelaajat eri klustereihin. Vaikka artikkeli ei itsessään pyrkinyt ennustamaan mitään pelaajien terveydestä tai kauden etenemisestä, voi sen tuloksista olla hyötyä pelaajien terveyteen liittyvien ennustemallien muuttu-

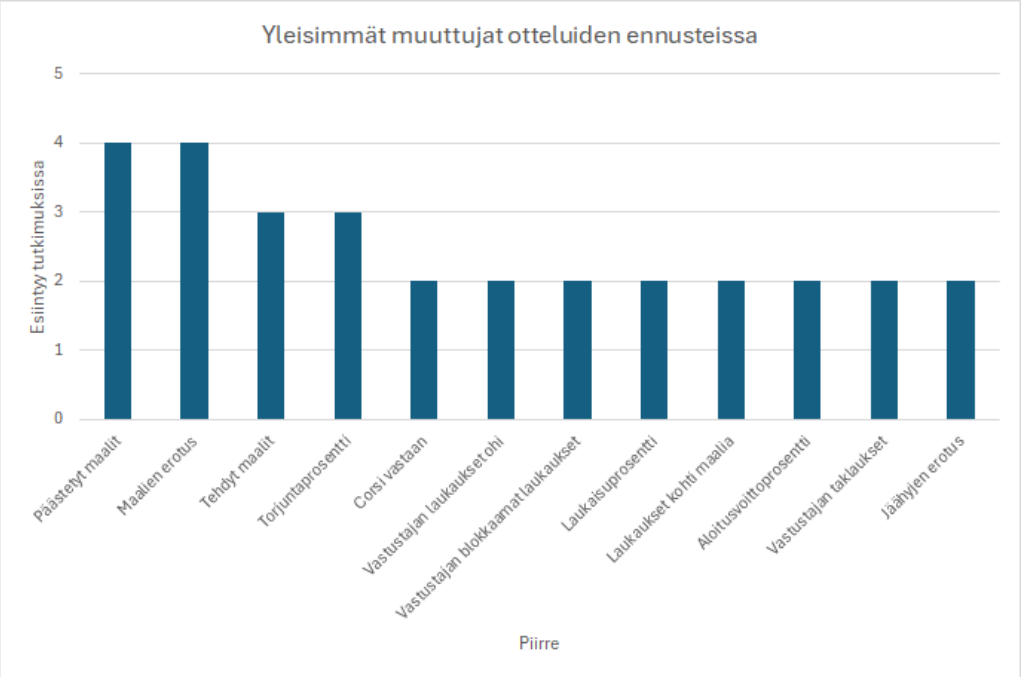
jia suunnitellessa tai esimerkiksi valmentajien suunnitellessa pelaajille henkilökohtaisia harjoituksia.

3.4 Yleisimmät mallit ja muuttujat ennusteissa

Jääkiekko-otteluista on tarjolla useita satoja tilastoja, joista monet toistuivat eri artikkeleissa. Myös koneoppimismallit ovat monessa tutkimuksessa samat ja etenkin tukivektorikone, satunnaismetsät, K-lähintä naapuria sekä naiivi Bayes -luokittelija korostuivat aiemmissa tutkimuksissa.

Tutkimuksissa oli useita erilaisia tapoja, miten oli päädytty valitsemaan koneoppimismalliin sopivat piirteet. Artikkelissa[35] piirteitä otettiin suoraan datasta, minkä lisäksi niistä muodostettiin itse frekvenssitilastoja, kuten summia ja keskiarvoja, sekä momentumpohjaisia tilastoja, kuten edellisen kolmen ottelun tilastoja ja niistä muodostettavia trendejä. Toiset tutkimukset avasivat todella tarkasti miten piirteet oli valittu, esimerkiksi artikkelissa[34] oli otteluista kerätty 18 eri piirrettä. Tämän jälkeen piirteille tehtiin pääkomponenttianalyysi (principal components analysis, PCA), jossa moniulotteista dataa yksinkertaistettiin vähempiulotteiseksi. Erilaisilla tilastollisilla testeillä selvitettiin vielä, mitkä muuttujista ovat merkittäviä. Tilastollista testiä käytettiin myös artikkelissa[37] rajaamalla datan 24 piirrettä lopulta 17 piirteeseen. Julkaisussa[39] datassa on kymmeniä muuttujia, joita ei ilmeisesti rajattu ollenkaan. Julkaisussa oli kuva muuttujien rakenteesta, mutta ei mainittu tehdäänkö muuttujille mitään. Julkaisussa[36] tuotiin esiin taulukossa 12 piirrettä, joita otteluista on kerätty. Tämän lisäksi näistä piirteistä muodostettiin vielä omia piirteitä, kuten keskiarvoja ja summia. Vaikka piirteet valittiin eri tavoin, monessa tutkimuksessa esiintyi kuitenkin samoja piirteitä (kuva 3.5).

Pelaajien terveyden ennustamiseen liittyviä tutkimuksia oli toteutettu melko vähän. Artikkelin[31] ennusteissa aivotärähdyksille sekä artikkelin[49] ennusteissa loukkaantumisille yhteisiä piirteitä olivat aiemmat loukkaantumiset sekä pelaajan ikä.



Kuva 3.5: Yleisimmät muuttujat otteluennusteissa

Loukkaantumisissa piirteiden rajaamiseen käytettiin varianssin kasvukerrointa (Variance Inflation Factor, VIF), jossa alle 10 saadut arvot hyväksyttiin piirteiksi. Pelaajien suorituksia ennustettaessa pelaajia pyrittiin luokittelemaan tutkimuksissa erilaisista näkökulmista, mutta usein samat tilastot vaikuttivat pelaajan luokitteluun. Julkaisu varattujen pelaajien todennäköisyydestä pelata otteluita NHL:ssä[45], kolmannen pelaajan lisäämisen vaikutus ketjun tilastoihin[46], neuroverkkojen hyödyntäminen pelaajien paremmuusjärjestyksessä GIM-mittarin avulla[40] sekä pelaajien luokitteleminen kolmeen eri luokkaan tilastojen ja videopelien avulla[41] korostivat piirteissä etenkin tehtyjä pisteitä, maaleja, torjuntaprosentteja sekä laukauksia. Maalivahtien osalta erilaisiin tasoluokkiin ryhmittely tehtiin torjuntaprosentin perusteella[42].

Etenkin kaikki maaleihin liittyvät tilastot korostuivat tutkimuksissa. Yleisesti voitaneen todeta, että otteluiden ennustamiseen liittyvät piirteet muodostettiin niistä tilastoista, mitkä olivat helposti saatavilla otteluista esimerkiksi suoraan liigojen internet-sivuilta tai muiden kokoamista tilastoista. Monissa tutkimuksissa oli toden-

näköisesti melko samat tilastot alunperin, mutta eri tutkimukset painottivat mallien piirteissä näistä eri muuttujia (taulukko 3.8).

Taulukko 3.8: Koneoppimismallien piirteiden valinta

	Ei valintaa	Piirteiden johtaminen	SelectBestK	PCA	KMO	Bartlett	Wilcoxon	VIF	Korrelaatio	Wrapper
Otteluennusteet										
Noel, J et al.[35]		X	X							
Gu W. et al.[34]				X	X	X	X		X	
Weissbock J. & Inkpen D.[4]		X								
Gu W. et al.[37]							X			
Chin J. et al.[39]	X									
Weissbock J et al.[36]	X									
Terveysennusteet										
Bruce J. et al.[31]	X									
Luu B. et al.[49]								X		
Pelaajaennusteet										
Brownlee S. et al.[46]									X	
Khan R. et al. [42]	X									
Liu G. & Schulte O. [40]		X								
Liu J. et al.[45]		X								
Lehmus T. et al.[41]									X	X

Neljä tutkimusta ei rajannut piirteitä ollenkaan. Neljä tutkimusta johti uusia piirteitä jo olevista piirteistä, esimerkiksi laskemalla summia ja keskiarvoja. Yksi tutkimuksista käytti SelectBestK-metodia, jolla tarkoitetaan sitä, että valittiin K-määrä parhaita piirteitä[35]. Näiden lisäksi käytettiin erilaisia tilastollisia testejä piirteiden rajaamiseen, varianssin kasvukerrointa, sekä piirteiden, jotka korreloivat toisten piirteiden kanssa poistamista. Yksi tutkimus käytti wrapper-menetelmää, joka tarkoittaa menetelmää, jossa pyritään löytämään ne piirteet, joilla mallit suoriutuvat parhaiten, kokeillen erilaisia piirteiden kombinaatioita [41].

4 Pohdinta

Suurin osa aiheeseen liittyvistä tutkimuksista kohdistui suurimpaan jääkiekkoliigaan NHL:ään, joten tutkimuksen tulokset voisivat olla hyvinkin erilaiset pienemmissä liigoissa. Samat lainalaisuudet ja satunnaisuus kuitenkin pätevät myös pienemmissä liigoissa[3], joten näitä koneoppimismalleja voi varmasti pyrkiä hyödyntämään myös muissakin jääkiekkoliigoissa.

Enemmistö tutkimuksista päätyi noin 60 % tarkkuuteen otteluennusteissa, kun ennustustarkkuuden ylärajaksi on aiemmin arvioitu noin 62 % tarkkuutta[4]. Artikkeli[37] saavutti 77,5 % tarkkuuden, mutta käytti ennusteissaan myös neljää asiantuntijaa, mikä saattaa lisätä mallin subjektiivisuutta. Sen lisäksi kyseisessä artikkelissa ennustettiin vain 89 pudotuspeliä. Pudotuspelit pelataan vasta runkosarjan jälkeen ja niissä ensimmäisillä kierroksilla runkosarjan parhaat joukkueet kohtaavat runkosarjan huonompia joukkueita, joten yli 62 % tarkkuus saatetaan saavuttaa vain ennustamalla paremmin runkosarjassa sijoittunutta joukkuetta. Suuri osa tutkimuksista oli tehty käyttämällä koneoppimismallien oletusparametreja, joten mallien tarkkuutta voitaisiin mahdollisesti parantaa optimoimalla parametreja. Artikkelit momentumin vaikutuksesta otteluennusteisiin[35], aivotärähdysten ennustamisesta[31], loukkaantumisten ennustamisesta[49], ketjumuutoksista[46], pelaajien luokittelusta GIM-mittarilla[40] sekä pelaajien luokittelusta tilastojen ja videopelien perusteella[41] olivat ainoat, joissa parametreja oli säädetty.

Terveyteen liittyvät mallit olivat melko tarkkoja jo julkisesti saatavilla tiedoil-

la, mutta niistä saataisiin todennäköisesti vielä tarkempia vain joukkueille saatavilla olevilla tiedoilla. Pelaajien luokittelussa mallit olivat myös tarkkoja, esimerkiksi neuroverkon ennustamat parhaat pelaajat saivat seuraavana kautena GIM-mittarin luokitteluun sopivat palkankorotukset[40]. Myös pelaajien luokittelu tilastojen ja videopelien perusteella[41] sekä maalivahtien luokittelu[42] osoittautuivat toimiviksi. Mallien tarkkuus voisi osittain selittyä sillä, että satunnaisuus saattaa vaikuttaa vähemmän yksittäisten pelaajien kohdalla kuin otteluissa kaikkine tapahtumineen.

Artikkeli[34] väittää saavuttaneensa parhaimmillaan 91,8% tarkkuuden otteluiden ennustamisessa. Tätä tulosta ei voi kiistää ilman tarkempaa datan tutkimista, mutta vaikuttaa mahdottomalta saada näin tarkkaa ennustetta jääkiekko-otteluille. Jos näin suuri tarkkuus saavutettaisiin, voisi vedonlyöjä tienata valtavaa taloudellista etua lyömällä vetoa mahdollisesti positivisella odotusarvolla[10] ja kaikki aiemmat tutkimukset jääkiekon hajonnasta ja 62 % ylärajasta osoittautuisivat epäluotettavaksi. Mallissa on saattanut tapahtua esimerkiksi datavuoto, joka antaa mallille oikeita vastauksia.

Vaikka artikkelin[34] 91,8 % kaltaiset tarkkuudet kuulostavat epärealistiselta ja viittaisivat datavuotoon tai ylioppimiseen, myös artikkelin[4] 62 % yläraja vaikuttaa melko matalalta. Jos pelkästään kotijoukkueen ennustaminen johtaa 56 % ennustustarkkuuteen[4], olisi ennustettava väli ainoastaan 56–62 %. Artikkeli [37] saavutti 77,5 % tarkkuuden 89 ottelulla, mutta myös artikkeli[39] saavutti 77,8 % tarkkuuden ennustaessa 5717 ottelua, mikä viittaisi siihen, että myös pidemmällä aikavälillä on mahdollista saavuttaa korkeampia tarkkuuslukuja. Tieteellisten tekstien ulkopuolelta Suomen jääkiekkoliigassa Digia Oyj saavutti 76 % tarkkuuden kauden 2023–2024 otteluennusteille käyttäen satunnaismetsää ennustamiseen[5]. Näin ollen noin 75 % yläraja kuulostaisi myös mahdolliselta ja etenkin 62 % kuulostaa melko alhaiselta. Toisaalta tämä voi viitata myös siihen, että eri tasoisten liigojen välillä voi olla myös eroja ennustustarkkuuden kannalta ja etenkin yksittäisten kausien välillä voi

olla eroja.

Suurin osa tutkimuksissa käytettävistä menetelmistä ovat nykypäivänä jo perinteisiä ja yleisessä käytössä olevia algoritmejä. Jääkiekko tarjoaa paljon aikasarjadataa, joten hieman kehittyneemmät algoritmit, kuten takaisinkytketyt neuroverkot, LSTM-verkko ja muuntimet voisivat ennustaa näitä perinteisiä malleja paremmin. Jääkiekon osalta ei löytynyt neuroverkkojen avulla tehtyjä ennusteita artikkelin[40] pelaajien paremmuusjärjestyksen sekä artikkelin[4] otteluennusteiden lisäksi. Muissa lajeissa, kuten jalkapallossa neuroverkkoja on yritetty hyödyntää, esimerkiksi artikkeli[52] saavutti LSTM-verkon avulla jopa 80,75 % tarkkuuden otteluiden ennustamisessa. Amerikkalaisessa jalkapallossa on hyödynnetty sekä konvoluutio- että takaisinkytkettyjä neuroverkkoja, joilla on saatu yli 99 % tarkkuuksia taklausennusteisiin[53]. Nämä tarkkuudet eivät tietysti ole suoraan verrattavissa jääkiekkoon, sillä lajeissa sattuman osuus on eri[3], mutta luultavasti tilastot ja ottelun kulku ovat sen verran lähellä toisiaan, että myös jääkiekkoon neuroverkkoja voitaisiin soveltaa. E-urheilussa puolestaan LSTM-algoritmi on saavuttanut jopa 95,59 % tarkkuuden käyttäen League of Legends -pelistä tarjolla ollutta aikasarjadataa[54]. Suurissa tarkkuuksissa kyseessä voi tietysti olla jälleen datavuoto tai ylioppiminen.

Suurin osa neuroverkkoihin ja urheilulajeihin liittyvistä tutkimuksista keskittyy ennustamisen sijaan tunnistamaan otteluista erilaisia tapahtumia, esimerkiksi koripalloilijan liikerataa takaisinkytketyllä neuroverkolla,[55], tennispelaajan liikerataa muuntimilla [56], sekä jalkapallotapahtumia, kuten maaleja, maaliyrityksiä, varoituksia ja kulmapotkuja sekä konvoluutio- että takaisinkytketyllä neuroverkolla[57]. Aikasarjadatan hyödyntämistä urheiluennusteita tehdessä on helpompi törmätä keskustelufoorumeihin siitä, miksi kukaan ei käytä aikasarjadataa urheiluennusteisiin [58], kuin aiheesta tehtyihin tutkimuksiin. Tulevaisuudessa jääkiekkoennusteiden ja muiden urheilulajien ennustamisen kannalta tutkimusta voitaisiinkin tehdä nimenomaan kehittyneempien aikasarja-algoritmien osalta.

5 Yhteenveto

Tässä tutkielmassa tarkkailtiin koneoppimismenetelmien hyödyntämistä jääkiekkodatan avulla tehtäviin ennusteisiin. Tutkielman tutkimuskysymykset olivat seuraavat:

1. Mitä koneoppimismenetelmiä on tähän mennessä hyödynnetty jääkiekkodatan avulla tehtäviin ottelu- ja pelaajaennusteisiin?

2. Kuinka tarkasti koneoppimismallit pystyivät tekemään jääkiekkodatan avulla ennusteita?

3. Mitkä tekijät vaikuttivat eniten mallin tekemään ennusteeseen?

Koneoppimisessa ei ole yhtä oikeaa algoritmia ratkaisun löytämiseen, ja sen huomaa myös tutkimusten useista eri algoritmeista. Tutkimuksissa esiin tuotiin K-keskiarvot, lineaarinen diskriminanttianalyysi, päätöspuut, XGBoost, neuroverkot, K-lähintä naapuria sekä etenkin useasti käytettiin logistista regressiota, tukivektorikonetta, satunnaismetsiä sekä naiivi Bayes-luokittelijaa.

Jääkiekossa tapahtumat tapahtuvat nopeasti ja satunnaisuudella on suuri vaikutus pelin kulkuun. Otteluita ennustaessa tarkkuudet olivat usein alle 60 %, mutta kaksi tutkimusta pääsi noin 77 % tarkkuuslukemiin ja yksi tutkimus saavutti 91,8 % tarkkuuden. Pelaajien seuraavan kauden loukkaantumista ennustaessa mallit osuivat todella tarkasti, noin 95 % tarkkuudella, ja aivotärähdyksiä ennustaessa AUC oli 0,79. Esimerkiksi aivotärähdyksissä tarkkuus on parempi kuin mitä NHL:n palka-

tuilla ottelutarkkailijoilla, joten koneoppimismallit ovat toimivia loukkaantumisten ennustamiseen. Myös pelaajien luokittelun kannalta koneoppiminen onnistui luokittelemaan pelaajia ja maalivahteja toimivasti.

Myös piirteitä oli useita erilaisia mallista riippuen ja joissakin malleissa niiden muuntamiseen ja karsimiseen perehdyttiin syvällisesti esimerkiksi tilastollisilla testeillä, kun toisissa tutkimuksissa kaikki mahdollinen data käytettiin mallin piirteinä. Otteluennusteissa datana voi yleisesti todeta käytettäneen dataa, jonka saa helposti ladattua erilaisilta internet-sivustoilta. Etenkin maaleihin ja laukaisuihin liittyvät tilastot korostuivat tutkimuksissa. Pelaajien terveyden kannalta ikä ja aiemmat loukkaantumistilastot olivat merkittäviä tekijöitä. Pelaajien suoriutumisen ja luokittelun kannalta tärkeimpiä muuttujia olivat usein erilaiset ottelusuoritukset, kuten tehdyt pisteet, maalit, torjuntaprosentit ja laukaukset.

Lähes jokainen tutkimus kohdistui suurimpaan jääkiekkoliigaan NHL:ään, joten jatkotutkimusten kannalta hyötyä voisi olla tutkia pienempiä, eri maiden liigoja, jotta aiemman kirjallisuuden yleistettävyys koko jääkiekkomaailmaan selkeytyisi. Aikaisemmissa tutkimuksissa käytetyt koneoppimismallit ovat nykypäivänä jo melko perinteisiä, joten esimerkiksi aikasarjadataan pohjautuvien neuroverkkojen, kuten takaisinkytketyn neuroverkon, LSTM-verkon sekä muuntimien tutkiminen voisi osoittautua hyödylliseksi.

Lähdeluettelo

[1] M. Honkanen, *Dataa & Kiekkoa, jakso 8: Runkosarjan digitaalit järjestykseen & ottelusarjojen ennakointia*, www.youtube.com/watch?v=86_xTLmRKX8&t=2m52s, viitattu 10.10.2024.

[2] MoneyPuck.com, https://moneypuck.com/g.htm?id=2024020254, viitattu 15.11.2024.

[3] M. J. Mauboussin, *The Success Equation: Untangling Skill and Luck in Business, Sports and Investing*. Harvard Business Review Press, 2012, s. 78–81.

[4] J. Weissbock ja D. Inkpen, "Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League", teoksessa *Advances in Artificial Intelligence. Canadian AI 2014*, M. Sokolova ja P. van Beek, toim., sarja Lecture Notes in Computer Science, vol. 8436, Springer, Cham, 2014. DOI: 10.1007/978-3-319-06483-3_22.

[5] T. Salmela, *Tekoäly ennustaa jääkiekon SM-liigan voittajan – näin malli on toteutettu*, https://digia.com/blogi/tekoaly-ennustaa-jaakiekon-sm-liigan-voittajan-nain-malli-on-toteutettu, viitattu 7.11.2024.

[6] S. Menon, *Why AI Is Popular Now – And Two Ways To Use It Better*, https://www.forbes.com/councils/forbestechcouncil/2023/12/01/why-ai-is-popular-now-and-two-ways-to-use-it-better/, viitattu 4.11.2024.

[7] K. Shailaja, B. Seetharamulu ja M. A. Jabbar, "Machine Learning in Healthcare: A Review", teoksessa *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, s. 910–914. DOI: 10.1109/ICECA.2018.8474918.

[8] M. Campbell ja S. Hong, "Predicting the outcome of sports competitions using Poisson distribution", *Issues in Information Systems*, vol. 25, nro 1, s. 188–198, 2024. DOI: 10.48009/1_iis_2024_116.

[9] L. Donaldson, B. Li ja M. D. Cusimano, "Economic burden of time lost due to injury in NHL hockey players", *Injury Prevention*, vol. 20, nro 6, s. 347–349, 2014.

[10] W. Edwards, "Probability-Preferences in Gambling", *The American Journal of Psychology*, vol. 66, nro 3, s. 349–364, 1953, ISSN: 00029556. url: http://www.jstor.org/stable/1418231 (viitattu 15. 11. 2024).

[11] T. Cover ja P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, nro 1, s. 21–27, 1967. DOI: 10.1109/TIT.1967.1053964.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations", 1967. url: https://api.semanticscholar.org/CorpusID:6278891.

[13] B. Mahesh, *Machine Learning Algorithms -A Review*, tammikuu 2019. DOI: 10.21275/ART20203995.

[14] J. Quinlan, "Introduction of decision trees", *Mach Learn*, vol. 1, s. 81–106, 1986.

[15] T. Chen ja C. Guestrin, "XGBoost: A Scalable Tree Boosting System", teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sarja KDD '16, San Francisco, Cali-

fornia, USA: Association for Computing Machinery, 2016, s. 785–794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. url: https://doi.org/10.1145/2939672.2939785.

[16] F.-J. Yang, "An Implementation of Naive Bayes Classifier", teoksessa *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, s. 301–306. DOI: 10.1109/CSCI46756.2018.00065.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988, ISBN: 1558604790.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[19] B. Cheng ja D. M. Titterington, "Neural Networks: A Review from a Statistical Perspective", *Statistical Science*, vol. 9, nro 1, s. 2–30, 1994. DOI: 10.1214/ss/1177010638. url: https://doi.org/10.1214/ss/1177010638.

[20] L. Alzubaidi, J. Zhang, A. J. Humaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions", *Journal of Big Data*, vol. 8, nro 1, s. 53, 2021. DOI: 10.1186/s40537-021-00444-8.

[21] M. I. Jordan, "Serial Order: A Parallel Distributed Processing Approach", Institute for Cognitive Science, University of California, San Diego, La Jolla, California, 92093, ICS 8604, 1986.

[22] Z. C. Lipton, J. Berkowitz ja C. Elkan, *A Critical Review of Recurrent Neural Networks for Sequence Learning*, 2015. arXiv: 1506.00019 [cs.LG]. url: https://arxiv.org/abs/1506.00019.

[23] S. Hochreiter ja J. Schmidhuber, "Long Short-term Memory", *Neural computation*, vol. 9, s. 1735–80, joulukuu 1997. DOI: 10.1162/neco.1997.9.8.1735.

[24] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need", teoksessa *Proceedings of the 31st International Conference on Neural Information Processing Systems*, sarja NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, s. 6000–6010, ISBN: 9781510860964.

[25] Y. Singhal, A. Jain, S. Batra, Y. Varshney ja M. Rathi, "Review of Bagging and Boosting Classification Performance on Unbalanced Binary Classification", teoksessa *2018 IEEE 8th International Advance Computing Conference (IACC)*, 2018, s. 338–343. DOI: 10.1109/IADCC.2018.8692138.

[26] X. Ying, "An Overview of Overfitting and its Solutions", *Journal of Physics: Conference Series*, vol. 1168, nro 2, s. 022 022, helmikuu 2019. DOI: 10.1088/1742-6596/1168/2/022022. url: https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

[27] S. Kaufman, S. Rosset, C. Perlich ja O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance", *ACM Trans. Knowl. Discov. Data*, vol. 6, nro 4, joulukuu 2012, ISSN: 1556-4681. DOI: 10.1145/2382577.2382579. url: https://doi.org/10.1145/2382577.2382579.

[28] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", *Mach. Learn. Technol.*, vol. 2, tammikuu 2008.

[29] M. B. A. McDermott, L. H. Hansen, H. Zhang, G. Angelotti ja J. Gallifant, *A Closer Look at AUROC and AUPRC under Class Imbalance*, 2024. arXiv: 2401.06091 [cs.LG]. url: https://arxiv.org/abs/2401.06091.

[30] Q. Wang, Y. Ma, K. Zhao ja Y. Tian, "A Comprehensive Survey of Loss Functions in Machine Learning", *Annals of Data Science*, vol. 9, nro 2, s. 187–212, huhtikuu 2022. DOI: 10.1007/s40745-020-00253-.

[31] J. Bruce, K. Riegrel ja W. M. et al., "A Machine Learning Approach to Concussion Risk Estimation Among Players Exhibiting Visible Signs in Professional Hockey.", *Sports Med*, 2024. DOI: https://doi.org/10.1007/s40279-024-02112-2. url: https://link.springer.com/article/10.1007/s40279-024-02112-2#citeas.

[32] S. Whyno, *Hockey business is booming as the NHL bounces back from the pandemic in a big way*, https://apnews.com/article/nhl-business-1c85057bf51c57a14b2bb9c9c781f9d6, viitattu 24.10.2024, 2024.

[33] International Ice Hockey Federation, *IIHF Official Rule Book 2024/25*, https://www.iihf.com/en/statichub/4719/rules-and-regulations, viitattu 24.10.2024.

[34] W. Gu, K. Foster, J. Shang ja L. Wei, "A game-predicting expert system using big data and machine learning", *Expert Systems with Applications*, vol. 130, s. 293–305, 2019, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.04.025. url: https://www.sciencedirect.com/science/article/pii/S0957417419302556.

[35] J. T. P. Noel, V. Prado da Fonseca ja A. Soares, "A Comprehensive Data Pipeline for Comparing the Effects of Momentum on Sports Leagues", *Data*, vol. 9, nro 2, 2024, ISSN: 2306-5729. DOI: 10.3390/data9020029. url: https://www.mdpi.com/2306-5729/9/2/29.

[36] J. Weissbock, H. L. Viktor ja D. Inkpen, "Use of Performance Metrics to Forecast Success in the National Hockey League", teoksessa *MLSA@PKDD/ECML*, 2013. url: https://api.semanticscholar.org/CorpusID:42267309.

[37] W. Gu, T. Saaty ja R. Whitaker, "Expert System for Ice Hockey Game Prediction: Data Mining With Human Judgment", *International Journal of Infor-*

mation Technology & Decision Making, vol. 15, kesäkuu 2016. DOI: 10.1142/S0219622016400022.

[38] T. L. Saaty, "The Analytic Network Process", teoksessa *Decision Making with the Analytic Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks*. Boston, MA: Springer US, 2006, s. 1–26, ISBN: 978-0-387-33987-0. DOI: 10.1007/0-387-33987-6_1. url: https://doi.org/10.1007/0-387-33987-6_1.

[39] J. S. Chin, F. H. Juwono, I. M. Chew, S. Sivakumar ja W. K. Wong, "Predicting Ice Hockey Results Using Machine Learning Techniques", teoksessa *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, 2023, s. 1–5. DOI: 10.1109/ICDATE58146.2023.10248726.

[40] G. Liu ja O. Schulte, "Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation", teoksessa *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, sarja IJCAI-2018, International Joint Conferences on Artificial Intelligence Organization, heinäkuu 2018, s. 3442–3448. DOI: 10.24963/ijcai.2018/478. url: http://dx.doi.org/10.24963/ijcai.2018/478.

[41] T. Lehmus Persson, H. Kozlica, N. Carlsson ja P. Lambrix, "Prediction of Tiers in the Ranking of Ice Hockey Players", teoksessa *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren ja A. Zimmermann, toim., Cham: Springer International Publishing, 2020, s. 89–100.

[42] R. Khan, P. Schena, K. Park ja E. Pinsky, "A Clustering Approach to Analyzing NHL Goaltenders' Performance", teoksessa marraskuu 2022, s. 3–10, ISBN: 978-3-031-17291-5. DOI: 10.1007/978-3-031-17292-2_1.

[43] National Hockey League, *NHL Goalie Stats 2023-2024*, https://www.nhl.com/stats/goalies?reportType=season&seasonFrom=20232024&seasonTo=20232024&gameType=2&sort=teamAbbrevs&page=0&pageSize=50, Viitattu 5.11.2024, 2023.

[44] PuckPedia, *Player Dashboard*, https://puckpedia.com/players/search, Viitattu 5.11.2024, 2023.

[45] O. Schulte, Y. Liu ja C. Li, "Model Trees for Identifying Exceptional Players in the NHL Draft", helmikuu 2018. DOI: 10.48550/arXiv.1802.08765.

[46] S. Brownlee, A. Khan, B. Vanderzyl ja M. El-Hajj, "Analysis of Hockey Forward Line Corsi: Should the Focus Be on Forward Pairs?", teoksessa *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, 2024, s. 0242–0248. DOI: 10.1109/CCWC60891.2024.10427925.

[47] National Hockey League, *Statistics*, https://www.nhl.com/stats/glossary, Viitattu 5.11.2024, 2024.

[48] NHL Public Relations, *NHL concussion evaluation and management protocol for 2022-23 season*, https://www.nhl.com/news/nhl-concussion-evaluation-and-management-protocol-for-2022-23-season-335002568, viitattu 25.10.2024, 2022.

[49] B. C. Luu, A. L. Wright ja H. S. e. a. Haeberle, "Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An analysis of 2322 Players From 2007 to 2017", *Orhopaedic Journal of Sports Medicine*, vol. 9, 2020. DOI: 10.1177/2325967120953404.

[50] C. Martin, A. Lieb, J. Tokish et al., "Initial versus Subsequent Injury and Illness and Temporal Trends Among Professional Hockey Players", *International Journal of Sports Physical Therapy*, vol. 19, nro 2, s. 215–226, helmikuu 2024. DOI: 10.26603/001c.92309.

[51] V. Rago, T. Fernandes ja M. Mohr, "Identifying Key Training Load and Intensity Indicators in Ice Hockey Using Unsupervised Machine Learning", *Research Quarterly for Exercise and Sport*, s. 1–13, 2024. DOI: 10.1080/02701367.2024.2360162.

[52] E. Tiwari, P. Sardar ja S. Jain, "Football Match Result Prediction Using Neural Networks and Deep Learning", teoksessa *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, s. 229–231. DOI: 10.1109/ICRITO48877.2020.9197811.

[53] D. Nimma ja A. Uddagiri, "Enhancing Tackle Prediction in NFL Games Through Feature Engineering and Hybrid Machine Learning Models", teoksessa *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2024, s. 1913–1919. DOI: 10.1109/I-SMAC61858.2024.10714680.

[54] C. Jung ja H. K. Kim, "Win Prediction from the Snowball Effect Perspectives", teoksessa *2022 IEEE Games, Entertainment, Media Conference (GEM)*, 2022, s. 1–6. DOI: 10.1109/GEM56474.2022.10017891.

[55] L. Liang, "Dynamic Tracking Optimization of Basketball Trajectory Based on Graph Neural Network", teoksessa *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, 2024, s. 197–201. DOI: 10.1109/ICSECE61636.2024.10729609.

[56] A. AlShami, T. Boult ja J. Kalita, "Pose2Trajectory: Using transformers on body pose to predict tennis player's trajectory", *Journal of Visual Communication and Image Representation*, vol. 97, s. 103 954, 2023, ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2023.103954. url: https://www.sciencedirect.com/science/article/pii/S1047320323002043.

[57] H. Jiang, Y. Lu ja J. Xue, "Automatic Soccer Video Event Detection Based on a Deep Neural Network Combined CNN and RNN", teoksessa *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2016, s. 490–494. DOI: 10.1109/ICTAI.2016.0081.

[58] u/jack7speed, *Why does no one use time series analysis for predictions?*, https://www.reddit.com/r/Sabermetrics/comments/s75fxa/why_does_no_one_use_time_series_analysis_for/, Viitattu 7.11.2024, 2022.