

Tuomas Mäkinen

URHEILUOTTELUIDEN TULOSTEN ENNUSTAMINEN KONEOPPIMISELLA

Kandidaatintyö
Informaatioteknologian ja viestinnän tiedekunta
Joulukuu 2024

TIIVISTELMÄ

Tuomas Mäkinen: Urheiluotteluiden tulosten ennustaminen koneoppimisella
Kandidaatintyö
Tampereen yliopisto
Tietotekniikan tutkinto-ohjelma
Joulukuu 2024

Urheiluotteluista kerätään suuria määriä dataa, jota voidaan hyödyntää koneoppimisen avulla. Koneoppimista voidaan soveltaa urheilussa monin eri tavoin, kuten loukkaantumisten ennustamiseen, tärkeiden suorituskyvyn attribuuttien tunnistamiseen sekä tulosten ennustamiseen, johon tämä työ keskittyy. Verkossa tapahtuvan vedonlyönnin kasvun myötä kiinnostus tulosten ennustamiseen on lisääntynyt.

Tämän työn tavoitteena on tutkia urheiluotteluiden tulosten ennustamista yleisten koneoppimismallien näkökulmasta. Tulosten ennustaminen on tyypillisesti binäärinen luokittelutehtävä, ellei lajissa ole tasapelin mahdollisuutta. Ennustamisessa keskeisiä osa-alueita ovat piirteiden keräys ja -valinta, sopivan koneoppimismallin valinta sekä mallin testaus.

Piirteiden valinta -algoritmien avulla parannettiin ennustustarkkuutta. Yleisten koneoppimismallien suoriutumisen vertailu on haastavaa, koska suorituskyky riippuu lajista ja tutkimuksessa käytetystä datasta. Mallien testauksessa data jaetaan opetus- ja testiaineistoon, yleensä k-ristinvalidoinnin tai segmentoinnin avulla. Ennustetarkkuus osoittautuu sopivaksi suorituskykymitariksi tulosten ennustamisessa.

Avainsanat: koneoppiminen, urheilu, tulosten ennustaminen

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

TEKOÄLYN KÄYTTÖ TÄSSÄ TYÖSSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

- Ei
 Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaprosessin aikana seuraavia tekoälysovelluksia:

Sovellus	Versio
ChatGPT	3.5

Tekoälyn käyttötarkoitus

Tässä työssä on käytetty tekoälyä, sellaisten termien kääntämisessä suomeksi, joille ei löytynyt suomennosta.

Osiot, joissa tekoälyä on käytetty

Osiossa 4 on käytetty tekoälyä termien suomennokseen.

Riskien tiedostaminen

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien osat, joiden tuottamisessa on hyödynnetty tekoälyä, ja hyväksyn vastuun mahdollisista tästä seuranneista eettisten ohjeiden rikkomuksista.

SISÄLLYSLUETTELO

1.	Johdanto	1
2.	Tutkimusmenetelmät	3
3.	Datan keräys ja käsittely	4
4.	Koneoppimismallit	6
4.1	Neuroverkot	6
4.2	Päätöspuut ja sattunaismetsät	7
4.3	Bayesin menetelmät	8
4.4	Tukivektorikone	9
5.	Mallin opetus ja testaus	10
6.	Lajien vertailu	12
6.1	Joukkuelajit	12
6.2	Yksilölajit	13
7.	Yhteenveto	14
	Lähteet	15

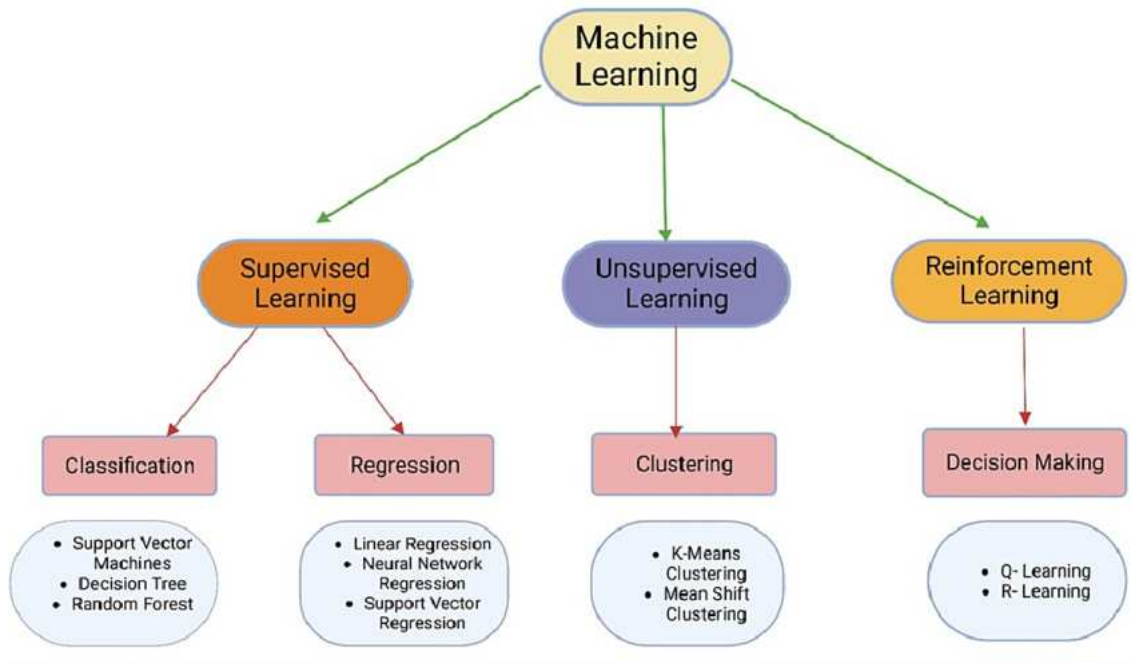
1. JOHDANTO

Nykypäivänä urheiluotteluista kerätään valtavia määriä dataa. Tätä dataa käytetään paljon data-analytiikassa, mutta lisäksi sitä voidaan käyttää tulevien otteluiden ennustamiseen. Urheiluotteluiden tulosten ennustamisesta on tullut mielenkiinnon kohde monille eri tahoille. Otteluiden ennustuksesta saatu informaatio on tärkeää joukkueiden ja pelaajien sidosryhmille, kuten osakkeenomistajille, sponsoreille, faneille, vedonlyöjille ja vedonvälittäjille. Verkossa tapahtuvan vedonlyönnin kasvun myötä otteluiden ennustuksessa kulkee suuria rahasummia. Vedonlyöjät haluavat maksimoida voittonsa ja vastakkaisesti vedonvälittäjät haluavat asettaa mahdollisimman tarkat vedonlyöntikertoimet otteluille. Joukkueen tai pelaajan valmentajat voivat käyttää ennustuksesta saatua informaatiota hyödykseen valmennuksessa.

Koneoppimisen yksi merkittävimmistä osa-alueista on ohjattu oppiminen 1.1. Ohjatun oppimisen yksi yleisimmistä tehtävistä on luokittelu. Luokittelussa tarkoitus on ennustaa luokka uudessa datassa. Ohjatussa oppimisessä tämä voidaan saavuttaa opettamalla luokittelija aikaisemmalla luokitetulla datalla. Urheiluotteluiden ennustamista käsitellään yleisesti luokitteluongelmana, jossa valitaan yksi luokka: voitto, häviö tai tasapeli (Bunker ja Thabtah 2019). Otteluiden ennustamista voi kuitenkin käsitellä ennustamalla ottelun pisteitä, mutta tämä ei ole niin yleistä. Ohjattu oppimien sopii hyvin urheiluotteluiden ennustamiseen, sillä luokiteltua dataa otteluista on helposti saatavilla.

Bunker ja Thabtah (2019) esittää kuusiosaisen rakenteen (SPR-CRISP-DM) otteluiden tulosten ennustamiselle. Askeet järjestyksessä ovat: alan ymmärtäminen, datan ymmärtäminen, datan valmistelu ja piirteiden valinta, koneoppimismallin valitseminen, mallin suorituskyvyn arviointi ja viimeinen askel on mallin käyttöönotto. Tämä rakenne laajentaa tiedonlouhinnassa käytettyä CRISP-DM -rakennetta urheiluotteluiden tulosten ennustuksissa yleisesti esiintyviin ongelmiin. Useat tulosten ennustamista käsittelevistä tutkimuksista seuraavat, joko CRISP-DM- tai SPR-CRISP-DM -rakenteita. Tässä työssä keskitytään enimmäkseen askeliin 3-5.

Työssä tarkoituksena on avata urheiluotteluiden tulosten ennustamisen tärkeimmät yhteiset seikat ja verrata tuloksia lajikohtaisesti. Työn näkökulma on rajattu muutamaaan kirjallisuudessa yleisimpään koneoppimismalliin, ja työssä ei paneuduta syvällisesti datan käsittelyyn tai alan ymmärtämiseen. Lyhyemmin tarkoitus on siis selvittää, kuinka hyvin



Kuva 1.1. Koneoppimisen osa-alueet

yleisimmät mallit suoriutuvat. Toisessa luvussa esitellään työn tutkimusmenetelmät. Kolmannessa luvussa käsitellään datan keräystä ja käsittelyä eli mistä ja miten data saadaan ja mitä toimenpiteitä pitää tehdä, ennen kuin sitä voidaan käyttää ennustuksissa. Neljännessä luvussa käsitellään koneoppimismallin valintaa. Luvun tarkoitus on kertoa mallien toimintaperiaatteista lyhyesti ja pohtia niiden soveltuvuutta urheiluotteluiden ennustukseen. Viidennessä luvussa kerrotaan yleisesti mallin validoinnista ja suorituskyvyn mittauksesta. Tämä yleensä sisältää valitun mallin testausta testidatalla ja mallin suorituskyvyn arvioimista mittaussuureilla, tarkkuudella. Kuudennessa luvussa verrataan työssä käsiteltyjen koneoppimismallien tuloksia lajikohtaisesti. Jokaiseen lajiin kuuluu paljon niiden omia yksityiskohtia, joten vertailu on hyvä tehdä lajikohtaisesti.

2. TUTKIMUSMENETELMÄT

Tämä kandidaatintyö on toteutettu kirjallisuuskatsauksena. Seuraavaksi käydään läpi tämän työn tutkimuskysymyksen rakentamiseen ja tiedonhakuun käytettyjä menetelmiä.

Työssä lähdettiin rakentamaan tutkimuskysymystä etsimällä ensin alustavia lähteitä. Tämä tehtiin Andor-hakupalvella. Alustavia lähteitä haettiin hakusanalla "machine learning AND sport AND predic*". Hakusanalla löytyi kaksi kirjallisuuskatsausta joukkuelajien tulosten ennustamisesta.

Tutkimuskysymyksen muodostamisen jälkeen työhön etsittiin lähteitä Tampereen yliopiston kirjaston ohjeistamista tietokannoista ja Andor:sta. Käytettyjä tietokantoja ovat ACM Digital Library, IEEE Electronic Library, Springer Link, ProQuest ja ScienceDirect. Tietokannoista ja Andor:sta haettiin lähteitä urheilulajikohtaisilla hakusanoilla. Esimerksi koripalloon liittyviä lähteitä haettiin hakusanalla "machine learning"AND (basketball OR NBA) AND (result* OR predic*). Samanlaisia hakusanoja käytettiin Tennistä, sulkapalloa ja jalkapalloa käsittelevien lähteiden etsinnässä.

Työssä yritettiin käyttää mahdollisimman uusia lähteitä. Suurin osa lähteistä on vuoden 2019 jälkeen. Kaikki urheiluotteluiden tuloksia käsittelevät tutkimukset ovat vertaisarvioituja. Tietoa koneoppimismalleista haettiin vertaisarvioimattomista lähteistä. Tämä ei kuitenkaan vaikuta työn luotettavuuteen, koska nämä tiedot ovat vakiintunutta alan tietoa.

Hakusanoilla hakemisen lisäksi lähteitä löytyi tietokantojen ehdotuksista. Esimerkiksi SpringerLinkissä on toiminto, joka ehdottaa samankaltaisia tutkimuksia sen perusteella, mitä muut lukijat ovat hakeneet luettuaan kyseisen lähteen. Tämän toiminon avulla löytyi joitain relevantteja lähteitä.

Suurin osa löytyneistä lähteistä sisältää lyhyen kirjallisuuskatsaus osion, missä kerrotaan aikaisemmista samaa lajia tai tekniikkaa käsitteleviä tutkimuksia. Näitä osioita hyödyntäen löytyi lisää hyödyllisiä lähteitä.

3. DATAN KERÄYS JA KÄSITTELY

Urheiluotteluista muodostuu valtavia datavirtoja, joita voidaan hyödyntää tulosten ennustamisessa. Tämän myötä urheiluotteluiden dataa keräävien tietokantojen määrä on myös kasvanut (Horvat ja Job 2020). Dataa kerätään usein julkisista tietokannoista tai esimerkiksi urheiluliigan virallisilta nettisivuilta. Osa tutkijoista on automatisoinut datan keräämisen kirjoittamalla ohjelmakoodia, joka hakee uuden datan verkosta (Bunker ja Thabtah 2019). Tätä menetelmää kutsutaan verkkoharavoimiseksi (web scraping).

Urheilutapahtumien ennustamiseksi malleille on annettava syöte, joka on saatavilla ennen tapahtuman alkua. Koska käytettävissä oleva data liittyy usein tapahtumien lopputuloksiin, kuten pisteisiin tai suoritusten määriin, sitä ei voi käyttää suoraan mallien opettamiseen. Siksi tietoja täytyy muokata. Ratkaisuna käytetään saatavilla olevien tietojen keskiarvoja, kuten joukkueen tai urheilijan keskimääräisiä piste- tai suoritusmääriä ennen tiettyä tapahtumaa. Jokaiselle tapahtumalle ja joukkueelle tai urheilijalle lasketaan keskiarvot seuraaville ominaisuuksille: pisteet, suoritukset, onnistuneet suoritukset, virheet, rangaistukset ja muut lajille ominaiset tilastot.

Muokattuja tietoja kutsutaan datan piirteiksi, ja ne ovat yksittäisiä ominaisuuksia tai muuttujia, jotka kuvaavat tarkasteltavaa kohdetta. Piirteet tarjoavat mallille olennaista tietoa ennusteen tekemiseksi. Urheiluotteluiden ennustamisessa piirteet voivat olla esimerkiksi joukkueiden aikaisemmat tulokset, maalimäärät, kulmapotkujen määrä, pelityyli tai pelaajakohtaiset tilastot, kuten loukkaantumiset ja pelikieltohistoria. Näitä piirteitä analysoimalla koneoppimismallit voivat tunnistaa yhteyksiä, jotka vaikuttavat ottelun lopputulokseen.

Piirteille tehdään usein jatkokäsittelyä, joita ovat piirteiden keräys ja valinta. Nämä ovat tärkeä alkuaskel koneoppimisessa. Piirteiden valinnassa tunnistetaan ja valitaan piirteitä datasta, joilla on suurin merkitys, ja hylätään piirteitä, joilla on vähiten merkitystä. Tällöin vähennetään tietojoukon ulottuvuutta. Piirteiden keräys on ulottuvuuden vähentämisen prosessi, missä alustava raakadata pienennetään helpommin käsiteltäviin joukkoihin. Piirteiden keräyksessä yhdistetään useita eri attribuutteja yhdeksi piirteeksi. Tällöin vähennetään prosessoitavan datan määrä kuitenkin säilyttäen alkuperäisen datan merkitys. Piirteiden keräys ja valinta tehostavat koneoppimismallien toimintaa, joka on erityisen tärkeää, kun käsitellään suuria tietojoukkoja. (Horvat ja Job 2020)

Piirteitä voidaan valita asiantuntijan osaamisen perusteella, käyttämällä piirteenvaal-

goritmeja tai yhdistämällä molempia menetelmiä. Varhaisissa tutkimuksissa tutkijat valitsivat piirteet vain oman lajin ymmärryksen mukaan, mikä usein heikensi mallin suoriutumista. Kuitenkin lähiaikoina on alettu käyttämään enemmän datan ohjaamia piirteenvalmamenetelmiä (Bunker ja Susnjak 2022). Sivistyneempiä piirteenvalmamenetelmiä käytetään maksimoimaan merkitykselliset piirteet ja minimoimaan tarpeettomat piirteet. Piirteenvalmamenetelmät voidaan yleensä jakaa filttereihin, wrappereihin, sisäisiin- ja hybridimenetelmiin (Jović, Brkić ja Bogunović 2015). Vain pieni määrä tutkimuksista hyödyntää asiantuntijan osaamista piirteiden valinnassa alustavien piirteiden valinnan ulkopuolella (Horvat ja Job 2020).

Useita erityyppisiä valinta-algoritmeja voidaan käyttää samanaikaisesti. (Sharma ym. 2022) käyttivät viittä eri piirteenvalmamenetelmää samanaikaisesti. Alkuperäisessä tietojoukossa oli 38 piirrettä, joka karsittiin 14 piirteeseen. Tämä tehtiin valitsemalla ainoastaan piirteet, jotka jokainen valintamenetelmä valitsi merkittäväksi piirteeksi. Tutkimuksessa parhaiten suoriutuva tietojoukko oli jokaisella menetelmällä karsittu tietojoukko.

4. KONEOPPIMISMALLIT

Työn näkökulma on rajattu muutamaaan kirjallisuudessa usein esiintyvään ohjatun oppimisen koneoppimismallin. Sopivien koneoppimismallien valinta on tärkeä osa otteluiden tulosten ennustamisen onnistumista. Melkein kaikki työssä käsitellyt tutkimukset käyttävät useita eri koneoppimismalleja ennustuksissa. Tämä on tärkeää, koska ei voida tietää, kuinka hyvin mallit suoriutuvat ilman vertailukohteita.

4.1 Neuroverkot

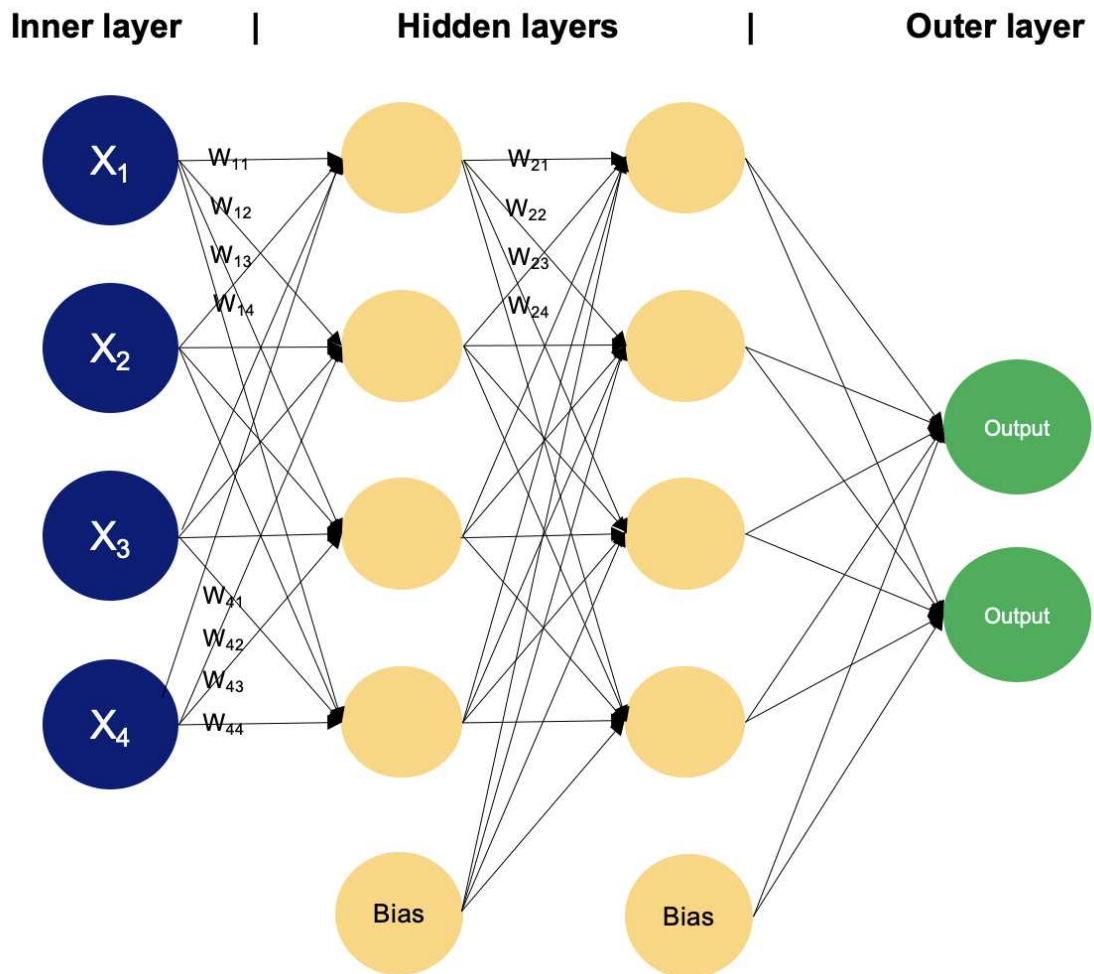
Neuroverkot ovat yksi käytetyimmistä koneoppimismalleista urheiluotteluiden tulosten ennustuksessa (Bunker ja Thabtah 2019). Neuroverkot toimivat matkimalla aivojen hermosolujen toimintaa. Ne koostuvat keinotekoisista "neuroneista", jotka ovat yhteydessä toisiinsa ja järjestetty eri kerroksiin. Jokainen neuroni vastaanottaa syötteitä muilta neuroneilta, muokkaa niitä painokertoimilla ja laskee yhteen saadakseen kokonaissyötteen. Tämän jälkeen syötesignaali kulkee aktivaatiofunktion läpi, joka määrittää, kuinka vahvana signaali välitetään seuraavalle tasolle. Yleinen neuroverkko tyyppi on kuvassa (4.1) esiintyvä multilayer perceptron eli MLP.

Verkko alkaa syötekerroksella, jossa data syötetään järjestelmään. Yksittäiseen neuroniin tulevaa informaatiota voidaan kuvata kaavalla 4.1,

$$z = \sum_{i=1}^n w_i x_i + b \quad (4.1)$$

jossa w_i on yhteyden painokerroin, x_i syöte ja b siirtymä. Kaikki neuroniin tulevat syötteet kerrotaan yhteyden painokertoimella, lisätään neuronin siirtymä ja summataan yhteen. Tämän jälkeen lasketaan aktivaatiofunktion arvo z :lla. Aktivaatiofunktio on usein epälineaarinen, jotta voidaan mallintaa epälineaarisia tehtäviä. Aktivaatiofunktioita ovat esimerkiksi sigmoidifunktio, hyperbolinen tangenti ja korjattu lineaarinen yksikkö (rectified linear unit).

Yleisin algoritmi, jolla neuroverkkoja opetetaan, on takaisinlevitys, jonka tarkoituksena on minimoida virhefunktio. Takaisinlevitys alkaa asettamalla satunnaiset arvot kaikille painokertoimille ja siirtymille. Tämän jälkeen lasketaan verkon ulostulo, josta lasketaan virhe.



Kuva 4.1. MLP

Koska tulosten ennustaminen on binäärinen luokittelu tehtävä, käytetään virheenä ristientropiaa. Kaikille parametreille lasketaan gradientti virheen suhteen ja siirretään parametreja gradientin vastaiseen suuntaan. Tämä operaatio toistetaan tietty määrä kertoja.

Neuroverkkojen opetus on työlästä variuksinkin, jos kyseessä on syvä neuroverkko, jolloin opetettavia parametreja on paljon. Lisäksi neuroverkoissa on paljon hienosäädettäviä hyperparametreja, kuten neuronien määrä kerroksessa, kerrosten määrä ja oppimisnopeus.

4.2 Päätöspuut ja sattunaismetsät

Päätöspuu on parametrin ohjatun oppimisen algoritmi, jota käytetään luokittelu ja regressio tehtävissä. Päätöspuu on hierarkkinen puurakenne, joka koostuu juurisolmista, haaroista, välisolmuista ja lehtisolmuista. Päätöspuut jakavat dataa ja tekevät päätöksiä toistuvasti kysymällä kysymyksiä datasta ja haarautumalla riippuen vastauksista. Näin ne pystyvät tekemään ennusteita tai luokituksia uusille datanäytteille.

Päätöspuilla on useampi yleinen opetusalgoritmi. Yksi näistä on Classification and Regres-

sion Tree (CART). Algoritmi jakaa opetusdatan kahteen osajoukkoon yhden piirteen ja raja-arvon mukaan. Piirre ja sen raja-arvo valitaan sen mukaan, millä arvoilla saadaan puhtaimmat osajoukot. Jaon jälkeen algoritmi toistaa aikaisemman toimenpiteen, kunnes saavutetaan maksimi syvyys.

Maksimisyvyys on parametri, joka määrittää kuinka monta kerrosta puussa on. Aikaisemmin sanottiin, että päätöspuut ovat parametrittomia. Tällä tarkoitetaan, että päätöspuut eivät tarvitse aikaisemmin määritettyjä parametreja toimiakseen jollain tasolla. Kuitenkin, jotta päätöspuut voisivat toimia hyvin, tarvitsee niiden muotoa ja kokoa rajoittaa. Tämä johtuu siitä, että rajoittamattomana päätöspuut sopeutuvat liika opetusdataan eli ylisovittaa. (Géron 2022)

Bunker ja Susnjak (2022) tutkimuksessa päätöspuut olivat toiseksi suosituin malli neuroverkkojen jälkeen. Syynä tälle on niiden nopea opetus ja se, että päätöspuissa ei ole hienosäädettäviä parametreja. Lisäksi päätöspuut eivät ole musta laatikko malleja, kuten neuroverkot, vaan ne ovat intuitiivisia ja helppoja ymmärtää. Tämä ominaisuus voi antaa eri tahoille informaatiota esimerkiksi siitä, mitkä piirteet ovat tärkeitä pelaajan tai joukkueen suoriutumiselle.

Pätöspuut ylisovittaa helposti ja niillä iso varianssi. Puun koon ja muodon rajoittamisen lisäksi on kehitetty menetelmä, jossa useiden päätöspuiden tuloksia käytetään valitsemaan lopullinen luokka. Tätä menetelmää kutsutaan sattunaismetsäksi. Ensin opetetaan monta päätöspuuta satunnaisilla piirteiden osajoukoilla. Kun päätöspuut ovat tehneet ennustukset, valitaan ennustuksista lopulliseksi luokaksi päätöspuiden moodi, eli yleisin luokka.

4.3 Bayesin menetelmät

Bayesin menetelmät perustuvat Bayesin teoreemaan. Tämä teoreema on perustavanlaatuinen konsepti todennäköisyyslaskennassa ja tilastotieteessä. Bayesin teoreemalla voidaan laskea ehdollinen todennäköisyys eli a posteriori. Yksinkertaistettuna tämä tarkoittaa tapahtuman A todennäköisyyttä, kun tapahtuma B tunnetaan. Oletetaan luokka muuttuja y ja ehdollinen piirre vektori x_n , voidaan teoreema yleistää luokitteluun kaavalla (4.2)

$$P(y|x_1\dots x_n) = \frac{P(y)P(x_1\dots x_n|y)}{P(x_1\dots x_n)} \quad (4.2)$$

Tämä yksinkertainen kaava on kaikkien Bayesin menetelmien pohja. Luokka voidaan valita maksimaalinen a posteriori (MAP) -päätöksellä. MAP-päätöksessä lasketaan luokille ehdollinen todennäköisyys kaikkien piirteiden suhteen ja valitaan luokka, jolla on suurin todennäköisyys.

Yksi yleinen joukko Bayesiläisiä menetelmiä on naiivit Bayes-luokittelijat (NB). Naiivi tässä tapauksessa tarkoittaa oletusta, että piirteet ovat keskenään ehdollisesti itsenäisiä. Tämä tarkoittaa, että piirteet eivät riipu toisistaan. NB on erittäin nopea opettaa ja vaatii huomattavasti vähemmän opetusdataa sivistyneempiin menetelmiin verrattuna (Pedregosa ym. 2011). Tämän takia NB on hyvä koneoppimismalli vertailukohdan asettamiseksi.

NB:n Ehdollinen itsenäisyys on vain oletus, joka ei usein pidä paikkaansa. Tämä aiheuttaa ongelmia luokittelussa. Lieventääkseen tätä ongelmaa, useita eri menetelmiä on ehdotettu. Yksi urheiluotteluiden tulosten ennustuksessa esiintyvä menetelmä on (Jiang ym. 2019) esittämä NB-CBFW (Naive Bayes Correlation Based Feature Weighting). Menetelmä kuuluu joukkoon menetelmiä, joissa piirteillä annetaan kerroin niiden tärkeyden mukaan. Menetelmät eroavat toisistaan siinä, että millä tavalla kertoimet lasketaan. NB-CBFW menetelmässä annetaan isompi kerroin piirteillä, joilla on suuri korrelaatio luokan kanssa ja mahdollisimman korreloimaton muiden piirteiden kanssa.

4.4 Tukivektorikone

Tukivektorikone eli SVM on tehokas ja monikäyttöinen koneoppimismalli, jolla voidaan tehdä lineaarista ja epälineaarista luokittelua ja regressiota. SVM toimii parhaiten pienissä ja keskikokoisissa tietojoukoissa eli noin sadoista alkioista tuhansiin (Géron 2022). Tämä sopii hyvin otteluiden tulosten ennustamiseen, koska urheiluotteluiden tietojoukkojen kokoa rajoittaa urheiluotteluiden määrä.

SVM perustuu siihen, että se pyrkii löytämään optimaalisen erotusrajan (hyperpinnan) luokkien välillä niin, että erotusraja on mahdollisimman kaukana lähimmistä tietopisteistä kummasakin luokassa. Tässä keskeinen idea on maksimoida luokkien välinen marginaali. Jos data on lineaarisesti eroteltavissa, SVM löytää suoran erotuspinnan luokkien välillä.

Kuitenkin urheiluotteluiden ennustus on harvemmin lineaarinen tehtävä. Jos data ei ole lineaarisesti eroteltavissa, SVM:ssä käytetään ydinfunktiota muuttamaan data suurempaan ulottuvuuteen. Yleinen ydinfunktio polynomiydin, jolla data muutetaan polynomiseen muotoon. Alhaisella polynomin asteella ei voi kuvata komplekseja tietojoukkoja ja korkealla polynomi asteella tulee laskettavia piirteitä todella paljon. Tähän ongelmaan löytyy kuitenkin ratkaisu nimeltään ydinkikka (kernel trick), jonka avulla polynomin asteita voidaan lisätä ilman, että piirteet kasvavat kombinatorisesti (Géron 2022). Muita ydinfunktioita on Gaussian RBF ja Sigmoidi. Ydinkikka toimii myös näissä ydinfunktioissa. Luokittelu usealle luokalle vaatii lisätyötä, mutta urheiluotteluiden ennustuksessa tämä ei monissa lajeissa aiheuta ongelmia.

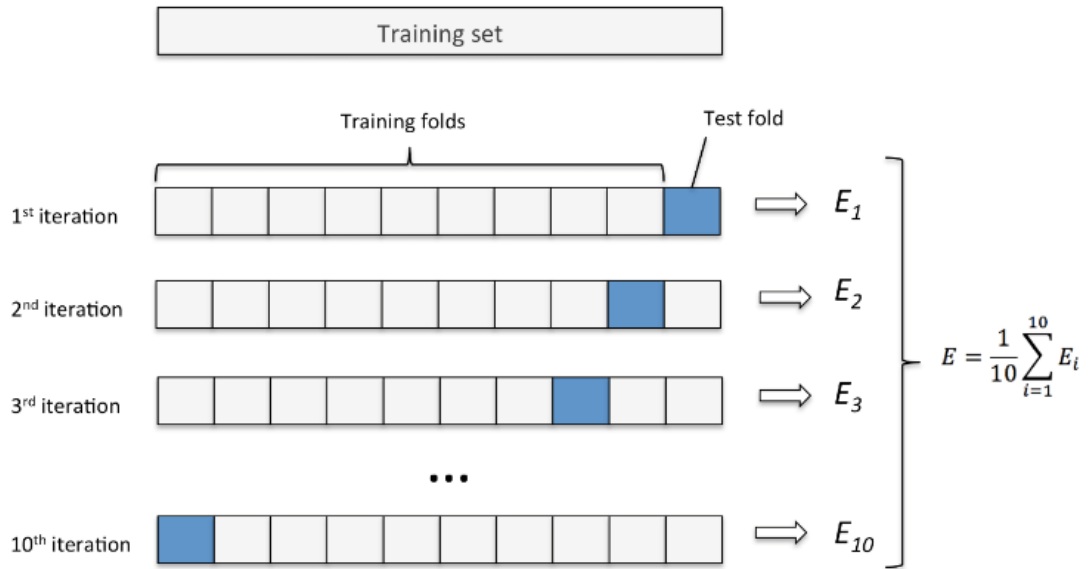
5. MALLIN OPETUS JA TESTAUS

Koneoppimistehtävissä data jaetaan usein opetus-, validointi- ja testausdataan. Opetusdatalla opetetaan malli, validointidataa käytetään mallin suorituskyvyn arvioimiseen ja hienosäätöön opetuksen aikana ja testausdatalla testataan malli vasta opetuksen jälkeen. Kun mallia koulutetaan ja testataan useilla eri datan osilla, voidaan arvioida, kuinka hyvin malli toimii uusilla tiedoilla. Tämä estää mallia oppimasta liikaa koulutusdatasta, mikä voisi johtaa heikompaan suorituskykyyn uusilla tiedoilla. Tilannetta, jossa malli oppii liikaa datasta ja ei yleisty hyvin uuteen dataan kutsutaan ylisovittamiseksi (overfitting).

Datan jakamiselle yleisesti kirjallisuudessa esiintyvä menetelmä on k-kertainen ristivalidointi, jossa data jaetaan k- yhtä suureen osaan (fold), ja jokaista osaa käytetään vuorollaan testiaineistona, kun taas loput osat toimivat koulutusaineistona. Yleinen arvo k:lle on 10, mikä on esitetty kuvassa (5.1). Menetelmän yleisyydestä huolimatta, sillä ei saavuteta parhaita tuloksia urheiluotteluiden ennustuksessa. Urheiluotteluiden tulosten ennustamisessa on tärkeää säilyttää otteluiden kronologinen järjestys opetusdatassa (Bunker ja Thabtah 2019), koska Urheilutapahtumat eivät ole täysin riippumattomia aikaisemmista tapahtumista (Horvat ja Job 2020). Tämän takia 10-kertainen ristiinvalidointi ei sovellu ilman asianmukaisia muutoksia, koska siinä sekoitetaan dataa satunnaisesti (Bunker ja Susnjak 2022).

Yleisesti paremmin suoriutuva menetelmä on datan segmentointi, missä data jaetaan yhdeksi opetusjaksoksi ja testausjaksoksi (Horvat ja Job 2020). Menetelmä suoriutuu usein paremmin, koska tällöin datan kronologisuus voidaan säilyttää helposti. Sopivan suhteen löytäminen opetusdatan ja testausdatan välillä riippuu tilanteesta, mutta yleisesti käytetty suhde koneoppimisessa on 80% opetusdataa ja 20% testausdataa. Jos käytetty malli sisältää hienosäädettäviä hyperparametreja, voidaan opetusdata vielä jakaa opetus- ja validointidataan. Opetus- ja validointidatan suhde on usein sama kuin opetus- ja testidatan välillä

Suorituskyvyn mittauksessa arvioidaan mallin kykyä suoriutua tehtävästään erilaisilla mittareilla. Työn tapauksessa arvioidaan mallin kykyä ennustaa otteluiden tuloksia. Käytetyt mittareita on useita, mutta ehkä yksi yleisimmin kirjallisuudessa esiintyvä mittari on tarkkuus. Tarkkuudella mitataan oikein ennustettujen tulosten suhdetta kaikkiin tuloksiin. Tulosten ennustamisessa. Tarkkuus on kohtuullinen suorituskyvyn mittari, koska se on



Kuva 5.1. 10-kertainen ristiinvalidointi

intuitiivinen, ymmärrettävä ja luokkien kokojen välillä ei todennäköisesti ole isoja eroja (Bunker ja Thabtah 2019), (Bunker ja Susnjak 2022). Kuitenkin suurin osa tutkimuksista käytti muita mittareita tarkkuuden lisäksi.

6. LAJIEN VERTAILU

Tulosten vertailussa yksinkertaisuuden vuoksi käsitellään suorituskyvyn mittarina vain tarkkuutta.

6.1 Joukkuelajit

Joukkuelajit ovat luontaisesti vaikeampia ennustaa, koska otteluissa on enemmän pelaajia. Jokaisen pelaajan suoriutuminen vaikuttaa tulokseen ja pelaajat voivat vaihdella ottelusta toiseen kokoonpanosta riippuen. Lisäksi joukkueen kokoonpano vaihtuu usein kausien välillä, jolloin kauden alussa ennustaminen on vaikeaa.

Iso osa koripalloa käsittelevistä tutkimuksista ennustaa NBA-otteluita. NBA (National Basketball Association) on Yhdysvalloissa ja Kanadassa toimiva miesten ammattilaiskoripalloliiga, joka on maailman tunnetuin ja arvostetuin koripalloliiga. Koripallo on ehkä helpoiten ennustettavissa oleva joukkuelaji, koska siinä on suuri pistemäärä, peliä pelataan sisällä ja siinä ei ole tasapelejä.

Thabtah, Zhang ja Abdelhamid (2019) tutkimuksessa ennustettiin NBA otteluiden tuloksia NB:llä, neuroverkoilla ja LMT:llä. LMT eli logistics model tree on päätöspuu, joka käyttää lehdissä logistista regressiota. Tutkimuksen tarkoitus on tunnistaa merkityksellisiä piirteitä käyttämällä useita piirteiden valinta algoritmeja ja tutkimalla mitkä tietojoukot suoriutuivat parhaiten. Täydellä tietojoukolla NB, LMT ja neuroverkot saavuttivat järjestyksessä 76%, 82% ja 83% tarkkuudet. RIPPER algoritmilla vähennetty tietojoukko nosti NB:n tarkkuutta 80%:n ja LMT:n tarkkuutta 83%:n, mutta neuroverkkojen tarkkuus laski 80%:n.

Toisin kuin muut työssä tarkastellut lajit jalkapallossa on mukana tasapelit, ja lajia pelataan ulkona. Tasapeli tapahtuu liigasta riippuen noin neljäsosassa peleistä. Tasapelin yleisyys luokkana vaikeuttaa tulosten ennustamista ja suorituskyvyn arviointia. Ennustamista vaikeuttaa lisäksi vähäinen pistemäärä ja vaihtelet sääolosuhteet.

Rodrigues ja Pinto (2022) ennustivat Englannin Valioliigan otteluita kahdeksalla mallilla. Tämä on enemmän malleja kuin yleensä käytetään. Työn näkökulmasta merkitykselliset mallit ovat NB, SVM, satunnaismetsät ja neuroverkot. Ensin malleja testattiin ilman piirteenvaihtamis-algoritmeja ja sitten valinta-algoritmien käytön jälkeen. Ensimmäisessä vaiheessa SVM suoriutui parhaiten saavuttaen 61,32% tarkkuuden. Valinta-algoritmien

käytön jälkeen, satunnaismetsät parhaiten suoriutuva malli, ja se saavutti 65,26%:n tarkkuuden.

6.2 Yksilölajit

Yksilölajeista on tehty vähemmän tutkimuksia. Tämä johtuu todennäköisesti niiden alhaisemmasta suosiosta. Yksilölajien pitäisi olla teoriassa helpompia ennustaa kuin joukkuelajien. Tämä johtuu siitä, että yksilölajeissa on vähemmän muuttuvia tekijöitä. Esimerkiksi yksilölajeissa ei tarvitse ottaa huomioon kokoonpanoon liittyviä tekijöitä.

Wilkins (2021) käsittelee tenniksen ennustamista vedonlyönnin näkökulmasta. Tutkimuksessa malleina ennustuksissa käytettiin logistista regressiota, neuroverkkoja, satunnaismetsiä, gradient boosting machine ja SVM:ää. Tarkkuus kaikilla käytetyillä malleilla olivat todella lähellä toisiaan. Tutkimuksessa saavutettiin noin 70% tarkkuus kaikilla malleilla. Tutkimuksessa määritettiin pelaajan sijoitus(ranking) pohjainen vertailumenetelmä, jolla saavutettiin 65% tarkkuus. Koneoppimismalleilla saavutettiin vain 5% korkeampi tarkkuus kuin mallittomalla vertailumenetelmällä. Tutkimuksessa käsitellyillä koneoppimismalleilla ja vedonlyönti strategiolla ei saavutettu pitkäaikaista nettovoittoa vedonlyönnistä.

Lei, Lin ja Cao (2024) ennustivat Australian avoimen tennisturnauksen, Wimbledonin tennisturnauksen miesten- ja naisten sarjan tuloksia. Ennustuksien luomiseen käytettiin SVM:ä ja kahta erityyppistä neuroverkkoa. Tutkimuksessa tärkeintä oli siinä käytetty opetusdata. Tutkimuksessa käytettiin vauhtia. Urheilun kontekstissa vauhdilla tarkoitetaan etua, jonka joukkue tai pelaaja saa, kun he ovat suoriutuneet lähiaikoina hyvin. Eli kun urheilija tai joukkue suoriutuu hyvin, tämä hyvä suoriutuminen kantautuu eteenpäin ottelun sisällä ja jopa tuleviin otteluihin. Vauhti on abstrakti käsite, jolla ei ole numeerista arvoa. Ensin pitää siis kvantifioida vauhti. Vauhti jaettiin kahteen osa-alueeseen: psykologinen vauhti ja strateginen vauhti. SVM:llä ja vauhdilla saavutettiin korkeimmillaan noin 84% tarkkuus Wimbledonin tennisturnauksessa ja 81.5% tarkkuus Australian avoimessa tennisturnauksessa. SVM:llä ja vauhdilla toimiva malli suoriutuu paremmin kuin muut mallit ilman vauhtia.

Sharma ym. (2022) ennustaa sulkapallo otteluita kolmella koneoppimismallilla: NB-CBFW, Composite Hypercubes on Iterated Random Projections (CHIRP) ja Hyper Pipes. Näistä parhaiten suoriutuva malli oli NB-CBFW. Menetelmä on suhteellisen uusi ja näyttää olevan lupaava urheiluotteluiden ennustamisessa. Kaksi jälkimmäistä mallia ei esiinny usein muissa tutkimuksissa eikä saavuttanut yhtä hyvää tarkkuuta verrattuna ensimmäiseen menetelmään. NB-CBFW mallilla saavutettiin joissakin tietojoukoissa 100 % tarkkuus. Sulkapallosta ei löytynyt enempää tutkimuksia kuin yksi.

7. YHTEENVETO

Urheiluotteluiden tulosten ennustamisesta koneoppimisella on tehty paljon tutkimuksia. Kuitenkin tutkimusten määrä on jakautunut kovin epätasaisesti lajien välillä. Jalkapallo ja koripallo ovat saaneet paljon huomiota tutkijoilta (Bunker ja Susnjak 2022). Tämä johtuu luultavasti siitä, että nämä lajit ovat maailman suosituimpia urheilulajeja.

Lajien välillä on suuria eroja ennustettavuudessa, mille on ehdotettu monia tekijöitä. Alhainen pistemäärä voi vaikeuttaa ennustamista. Lajin ja liigan kilpailun taso vaikuttaa ennustettavuuteen, koska alhaisen kilpailun lajissa tai liigassa on yleensä vain muutama ylivoimainen joukkue tai pelaaja. Tasapelit vaikeuttavat ennustamista. Vaikka vaikuttavia tekijöitä on tunnistettu, lajien luontaisen ennustettavuuden määrittäminen on edelleen hankalaa. (Bunker ja Susnjak 2022)

Tutkimukset käyttävät usein erilaisia piirteitä ja erilaisia piirteiden valintamenetelmiä, mikä vaikeuttaa koneoppimismallien välistä vertailua. Kuitenkin voidaan todeta, että SVM on usein parhaiten suoriutuvien mallien joukossa riippumatta syötteestä. Päättöpuut ja satunnaismetsät suoriutuvat hyvin vähennetyissä tietojoukoissa. Vastoin kuin muut työssä käsitellyt mallit, neuroverkot eivät näytä hyötyvän piirteidenvalinnasta yhtä paljon. Joissakin tapauksissa neuroverkot suoriutuvat huominkin vähennetyille piirteillä. NB ei useimmiten pärjää yhtä hyvin kuin muut mallit, mutta se on nopea opettaa, mikä tarjoaa helpon vertailukohdan. NB-CBFW suoriutuu hyvin sulkapallossa, mutta tutkimuksia muista lajeista ei löytynyt.

Tämä kandidaatintyö on rajoitettu vain muutamiin koneoppimismalleihin. Lähteissä esiintyi useita hyvin suoriutuvia malleja, jotka olivat tämän rajoitteen ulkopuolella, ja täten jäivät käsittelemättömäksi työssä. Vauhdin lisäksi työssä ei käsitelty lajin mallinnusta syvällisellä tasolla. Joissakin tutkimuksissa esiintyy menetelmiä, joissa piirrejoukkoa muokataan kuvaamaan lajia paremmin. Tämä vaati syvällisempää ymmärrystä ennustettavasta urheilulajista. Näillä sivistyneemmillä piirrejoukoilla voidaan mahdollisesti saavuttaa parempia tuloksia.

LÄHTEET

- Bunker, Rory ja Teo Susnjak (2022). "The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review". *Journal of artificial intelligence research*. DOI: <https://doi.org/10.1613/jair.1.13509>.
- Bunker, Rory ja Fadi Thabtah (2019). "A machine learning framework for sport result prediction". *Applied Computing and Informatics*. DOI: <https://doi.org/10.1016/>.
- Géron, Aurélien (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. eng. 3. painos. Sebastopol: O'Reilly Media, Incorporated.
- Horvat, Tomislav ja Josip Job (2020). "The use of machine learning in sport outcome prediction: A review". *WIREs Data Mining and Knowledge Discovery*. DOI: <https://doi.org/10.1002/widm.1380>.
- Jiang, Liangxiao ym. (2019). "A Correlation-Based Feature Weighting Filter for Naive Bayes". *IEEE Transactions on Knowledge and Data Engineering* 31.2, ss. 201–213. DOI: 10.1109/TKDE.2018.2836440.
- Jović, A., K. Brkić ja N. Bogunović (2015). "A review of feature selection methods with applications". *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, ss. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.
- Lei, Yilin, Ao Lin ja Jianuo Cao (2024). "Rhythms of Victory: Predicting Professional Tennis Matches Using Machine Learning". *IEEE Access* 12, ss. 113608–113617. DOI: 10.1109/ACCESS.2024.3444031.
- Pedregosa, F. ym. (2011). "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12, ss. 2825–2830.
- Rodrigues, Fátima ja Ângelo Pinto (2022). "Prediction of football match results with Machine Learning". *Procedia Computer Science* 204. International Conference on Industry Sciences and Computer Science Innovation, ss. 463–470. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2022.08.057>.
- Sharma, Manoj ym. (2022). "Naive bayes-correlation based feature weighting technique for sports match result prediction". eng. *Evolutionary intelligence* 15.3, ss. 2171–2186. DOI: <https://doi.org/10.1007/s12065-021-00629-3>.
- Thabtah, Fadi, Li Zhang ja Neda Abdelhamid (2019). "NBA Game Result Prediction Using Feature Analysis and Machine Learning". *Annals of Data Science*. DOI: <https://doi.org/10.1007/s40745-018-00189-x>.

Wilkins, Sascha (2021). "Sports prediction and betting models in the machine learning age: The case of tennis". *Journal of sports analytics*, ss. 99–117. DOI: <https://doi.org/10.3233/JSA-200463>.