Aalto-yliopisto
Aalto-universitetet
Aalto University

**Insinööritieteiden kandidaattiohjelma**

# Koneoppimismenetelmät alueiden saavutettavuustutkimuksessa

**Janne Toppi**

**Kandidaatintutkielma**
**2025**

Copyright ©2025 Janne Toppi

**Tekijä** Janne Toppi

**Työn nimi** Koneoppimismenetelmät alueiden saavutettavuustutkimuksessa

**Koulutusohjelma** Insinööritieteiden kandidaattiohjelma

**Pääaine** Geoinformatiikka

**Valvoja** Vanhempi yliopiston lehtori Petri Rönnholm

**Työn ohjaaja(t)** Apulaisprofessori Henrikki Tenkanen

**Päivämäärä** 08.12.2025 **Sivumäärä** 33 **Kieli** Suomi

## Tiivistelmä

Tässä kandidaatintyössä tarkastellaan kirjallisuustutkimuksen avulla, miten ja millaisia koneoppimismenetelmiä on sovellettu alueiden saavutettavuustutkimuksessa, millaisia mahdollisuuksia ja haasteita niiden käyttöön liittyy ja millaisia ongelmia niillä on yritetty ratkaista. Työ rajautuu saavutettavuuden mittaamiseen palveluiden ja kaupunkien välisen saavutettavuuden konteksteissa. Teoreettisena taustana esitellään koneoppimista ja tässä kandidaatin työssä esiintyviä keskeisiä algoritmeja sekä saavutettavuutta käsitteenä ja saavutettavuusmittareita, kuten HAM ja SIM.

Kirjallisuustutkimus osoittaa, että sekä ohjattua oppimista että ohjaamatonta oppimista voidaan menestyksekkäästi hyödyntää saavutettavuuden mallintamisessa. Ohjaamattomiin menetelmiin perustuva klusterointi soveltuu palveluiden saavutettavuusluokkien (esim ”hyvä”–”heikko”) tunnistamiseen silloin, kun valmiita luokkia ei ole. Hybridimallit (ohjaamattoman ja ohjatun oppimisen yhdistelmä) joissa klustereita käytetään ohjatun oppimisen koulutusdatana, mahdollistavat saavutettavuuden automaattisen luokittelun kaupunkialueilla. Gravitaatiopohjaisten saavutettavuusmittareiden ja imputoidun LASSO-regression yhdistäminen puolestaa auttaa tunnistamaan sosiodemografisia ja alueellisia tekijöitä, jotka selittävät saavutettavuuden alueellista vaihtelua. Lisäksi satunnaismetsään perustuvat mallit pystyvät toistamaan Hansenin saavutettavuusmallin käyttäytymistä ja ennustamaan saavutettavuutta tilanteissa, joissa virtausdata on puutteellista.

Tulosten perusteella koneoppimismenetelmät tarjoavat lupaavan lähestymistavan saavutettavuuden mallintamiseen, erityisesti kun käsitellään suuria ja monimuotoisia paikkatietoaineistoja. Keskeisiä haasteita ovat muun muassa datan laatu ja saatavuus, mallien tulkittavuus ja ylioppimis- ja alioppimisongelmat.

**Avainsanat** koneoppiminen, saavutettavuus, saavutettavuustutkimus, satunnaismetsä, Hansenin saavutettavuusmalli, spatiaalinen interaktiomalli, LASSO-regressio

**Author** Janne Toppi

**Title of thesis** Machine Learning Methods in Spatial Accessibility Research

**Programme** Bachelor's Programme in Engineering

**Major** Geoinformatics

**Thesis supervisor** Senior University Lecturer Petri Rönnholm

**Thesis advisor(s)** Assistant Professor Henrikki Tenkanen

**Date** 08.12.2025 **Number of pages** 33 **Language** Finnish

**Abstract**

In this bachelor's thesis, a literature review is used to examine how and what kinds of machine learning methods have been applied in spatial accessibility research, what opportunities and challenges are associated with their use, and what types of problems these methods have been used to address. The work focuses on measuring accessibility in the context of services and inter-urban accessibility. The theoretical backgroud introduces machine learning as a concept and the key algorithms used in this theses and the concept of accessibility and accessibility measures such as HAM and SIM.

The literature reviwe shows that both supervised and unsupervised learning can be successfully applied to accessibility modelling. Clustering algorithms based on unsupervised learning are suitable for identifying accessibility classes (e.g., "good" –"bad") in situations where predefined classes do not exist. Hybrid models – combining supervised and unsupervisen learning, where clusters are used as training data for supervised models – enable the automatic classification of accessibility in urban areas. Combining gravity-based accessibility measures with imputed LASSO regression helps identify sociodemographic and regional factors that explain spatial variation in accessibility. Additionally, random forest based models are capable of replicating the behaviour of Hansen's accessibility model and predicting accessibility in situations where flow data are inclomplete.

Based on results, machine learning mehtods provide a promising approach to modelling accessibility, particularly when working with large and heterogeneous spatial datasets. However, key challenges include the quality of the data and its availability, model interpretability and issues relatted ot overfitting and underfitting.

**Keywords** machine learning, spatial accessibility, accessibility modelling, random forest, Hansen's accessibility model, spatial interaction model, LASSO regression

# Sisällys

Operaattorit ja lyhenteet ... 6
Operaattorit ... 6
Lyhenteet ... 6
1 Johdanto ... 7
2 Koneoppimisen teoreettinen tausta ... 8
2.1 Ohjattu oppiminen ... 8
2.1.1 Regressio ... 8
2.1.2 Luokittelu ... 9
2.2 Ohjaamaton oppiminen ... 9
2.2.1 Klusterointi ... 9
2.2.2 Dimension vähentäminen ... 10
2.3 Algoritmit ... 10
2.3.1 K-Nearest Neighbors ... 10
2.3.2 K-Means ... 11
2.3.3 Päätöspuu ja satunnaismetsä ... 11
2.3.4 Neuroverkko ... 12
3 Saavutettavuustutkimuksen teoreettinen tausta ... 13
3.1 Saavutettavuuden komponentit ... 13
3.2 Saavutettavuuden mittarit ... 14
3.3 Spatiaalinen interaktiomalli ... 15
3.4 Hansenin saavutettavuusmalli ... 16
4 Koneoppiminen saavutettavuustutkimuksessa ... 17
4.1 Koneoppimismenetelmien käyttökohteet saavutettavuudessa ... 17
4.1.1 Koneoppimisen soveltaminen palveluiden saavutettavuuteen ... 17
4.2 Koneoppimisen yhdistäminen saavutettavuusteorioihin ... 26
5 Johtopäätökset ... 30
5.1 Tulokset ja havaintoja kirjallisuudesta ... 30
5.2 Haasteet ja kehityssuunnat ... 31
5.3 Potentiaali käytännön sovelluksilla ... 31
Lähteet ... 32

# Operaattorit ja lyhenteet

## Operaattorit

| | |
|---|---|
| $\Sigma_i$ | Summa indeksin *i yli* |
| max(x,y) | Suurempi arvo x ja y välillä |

## Lyhenteet

| | |
|---|---|
| Adam | Adaptive Moment Estimation |
| CART | Classification and Regression Tree |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| HAM | Hanses Accessibility Model |
| KNN | K-Nearest Neighbours |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LDA | Linear Discriminant Analysis |
| LR | Linear Regression |
| ML | Machine Learning |
| MLP | Multilayer Perceptron |
| NB | Naive Bayes |
| NN | Neural Network |
| OLS | Ordinary Least Squares |
| ReLU | Rectified Linear Unit |
| RF | Random Forest |
| SIM | Spatial Interaction Model |
| SVM | Support Vector Machine |

# 1 Johdanto

Koneoppimismenetelmät avaavat uusia mahdollisuuksia alueiden saavutettavuutta tarkastelevaan tutkimukseen, jossa pyritään ymmärtämään, kuinka helposti ihmiset, palvelut ja toiminnot ovat saavutettavissa eri sijainneissa. Saavutettavuutta on tutkittu jo monia vuosikymmeniä ja se on saanut monenlaisia eri määritelmiä. Yksi niistä – ja ehkäpä kuuluisin – on mahdollisuus, joka yksilöllä tietyssä sijainnissa on osallistua tiettyyn toimintaan tai joukkoon toimintoja (Hansen, 1959). Tällä tarkoitetaan yksinkertaistettuna sitä, kuinka helposti ihmiset, palvelut ja toiminnot ovat saavutettavissa tietyistä sijainneista.

Saavutettavuustutkimus on keskeinen osa yhdyskuntasuunittelua, kaupunkisuunittelua ja liikennejärjestelmien kehittämistä. Sen merkitys on olennainen yksilön arjen sekä yhteiskunnan toimivuuden kannalta. Perinteisesti saavutettavuutta on mitattu muun muassa matka-aikojen, fyysisten etäisyyksien ja liikennekustannusten perusteella. Hansenin saavutettavuusmalli (engl. Hansen's accessibility model, HAM) on eräs näistä. Nykypäivän koneoppimismenetelmien käyttöä saavutettavuuden mallintamiseen on tutkittu käyttäen esimerkiksi HAM:ia, mutta aihe on vielä varsin tuore varsinkin Pohjoismaissa.

Koneoppiminen on tekoälyn osa-alue, jossa tietokoneohjelmat oppivat tekemään päätelmiä tai ennusteita datasta ilman, että niitä on erikseen ohjelmoitu kaikkia sääntöjä varten. Koneoppimismenetelmiä voidaan jaotella useisiin ryhmiin sen perusteella, minkälaisia tehtäviä ne ratkaisevat. Eräitä koneoppimisen osa-alueita ovat valvottu oppiminen ja valvomaton oppiminen. Nämä ovatkin juuri niitä koneoppimisen osa-alueita, joita on käytetty saavutettavuustutkimuksessa.

Tässä kandidaatintyössä tutkitaan kirjallisuustutkimuksen avulla minkälaisia koneoppimismenetelmiä on sovellettu saavutettavuuden tutkimukseen ja minkälaisia mahdollisuuksia ja haasteita niiden käyttöön liittyy sekä minkälaisia ongelmia niillä on yritetty ratkaista tai niillä on mahdollista ratkaista. Toisessa luvussa tutustutaan koneoppimisen teoreettiseen taustaan, sekä algoritmeihin, joita on käytetty saavutettavuustutkimuksessa. Kolmannessa luvussa tutustutaan saavutettavuustutkimukseen sekä sen yleisiin mittareihin. Neljännessä luvussa tutustutaan siihen, kuinka koneoppimista on käytetty saavutettavuustutkimuksessa ja lopuksi tarkastellaan tutkimusten ja tutkimuskysymysten tuloksia sekä pohditaan tutkimuksen johtopäätöksiä.

# 2 Koneoppimisen teoreettinen tausta

Koneoppiminen on joukko menetelmiä, joiden tarkoituksena on tehdä päätelmiä, ennusteita tai havaintoja datasta. Nämä menetelmät ovat datapohjaisia menetelmiä, jotka yhdistävät tietojenkäsittelytieteen perustavanlaatuiset teoriat tilastotieteen, todennäköisyysteorian ja optimoinnin teorioiden kanssa. (Mohri ym., 2018.) Koneoppimismenetelmät voidaan jakaa erilaisiin oppimistyyleihin riipuen siitä minkälaista dataa käsitellään ja mikä tavoite datankäsittelyllä on.

## 2.1 Ohjattu oppiminen

Ohjattu oppiminen (engl. supervised learning) on oppimista, jossa oppija saa joukon valmiiksi luokiteltuja (eli merkittyjä) esimerkkitapauksia koulutusdatana ja tekee ennusteita kaikille ennennäkemättömille datapisteille (Mohri ym., 2018). Ohjattua oppimista siis käytetään, kun halutaan saada jokin ennuste. Ohjatun oppimisen malli täytyy ensin kouluttaa ja testata olemassa olevalla datalla, jonka jälkeen se optimoidaan niin, että se tuottaa luotettavia tai hyviä ennusteita. Ohjatun oppimisen algoritmeja on monia. Tässä alaluvussa tutustutaan ohjatun oppimisen eri tyyppeihin.

### 2.1.1 Regressio

Regressio on tekniikka ohjatussa oppimisessa, jossa tavoitteena on mallintaa riippuvuussuhde yhden tai useamman selitettävän muuttujan x ja tavoitemuuttujan y välillä. Toisin sanoen yritetään ennustaa reaalilukuarvo (jatkuva), kun saadaan jokin syöte. Kumarin ja Bhatnagarin (2022) mukaan regressiota on pääosin kuudenlaista: Lineaarista, Logistista, Polynomista, Ridge, Bayes lineaarista ja LASSO. Lineaarista regressiota voi käyttää, kun riippuvan ja selitettävien muuttujien välillä on likimain lineaarinen yhteys. Polynominen regressio on käytännössä lineaarisen regression laajennus, jossa malliin lisätään korkeampia potenssitermejä. Tämä regressiotyyppi sopii tilanteeseen, jossa muuttujien välinen suhde on epälineaarinen, mutta jatkuva. Logistinen regressio ei ennusta jatkuvaa arvoa vaan ennustaa kuuluvatko havainnot johonkin tiettyyn luokkaan esimerkiksi 1 tai 0, kyllä tai ei. Ridge (L2 säännöstys) on sopiva malli silloin, kun mallissa esiintyy multikollineaarisuutta, eli selittävät muuttujat korreloivat keskenään. Ridge lisää rangaistustermin, joka pienentää kertoimia ja vähentää mallin herkkyyttä datan vaihtelulle. Lasso (L1 säännöstys) toimii kuten Ridge, mutta rangaistustermi perustuu L1 normiin eikä L2 normiin kuten Ridgessä.

### 2.1.2 Luokittelu

Luokittelu on ohjatun oppimisen menetelmä, jossa tavoitteena on yrittää ennustaa jokin kategorinen tavoitemuuttuja eli määrittää, mihin luokkaan tai ryhmään kukin havainto kuuluu. Luokittelussa – toisin kuin regressiossa — ennustetaan diskreettiä luokkaa kuten ”hyvä” tai ”huono” saavutettavuus tai ”korkea” tai ”matala” palvelutaso. Binääriluokittelu on luokittelua, jossa malli ennustaa kahden eri luokan välillä, kuten edellä olevassa esimerkissä. Moniluokkaluokittelussa (engl. Multiclass calssification) malli ennustaa enemmän kuin kahden luokan välillä. Tällöin luokat voivat olla esimerkiksi erilaisia eläimiä, kuten koira, kissa tai elefantti. Yleisimpiin luokittelualgoritmeihin kuuluu päätöspuut (DT), jossa mallilla on puumainen rakenne, missä jokainen solmu (engl. node) kuvaa testiä jollekkin attribuutille ja jokainen oksa (engl. branch) on testin tulos. Lehtisolmu on tällöin lopullinen luokka. Satunnaismetsä on kokonaisuusmetodi (engl. ensemble method), joka käyttää useaa päätöspuuta parantaakseen tarkkuutta. KNN on algoritmi, joka luokittelee datapisteet sen k:n lähimmän naapurin enemmistöluokan perusteella. Tukivektorikone on taas malli, joka etsii optimaalisen hypertason (engl. hyperplane), joka erottaa parhaiten eri luokat toisistaan.

## 2.2 Ohjaamaton oppiminen

Ohjaamaton oppiminen (engl. unsupervised learning) on oppimista, jossa oppija saa pelkästään merkitsemätöntä koulutusdataa ja tekee ennusteita kaikille ennennäkemättömille datapisteille (Mohri ym., 2018). Ohjatun oppimisen tavoite on löytää datasta piilossa olevia kuvioita, rakenteita tai yhteyksiä ja suhteita. Se eroaa ohjatusta oppimisesta siten, että data ei sisällä valmiita vastauksia. Tavoitteena siis ei ole ennustaa tuntemattomia arvoja, vaan paljastaa datan sisäisiä rakenteita. Tämä tapahtuu esimerkiksi ryhmittelemällä dataa klustereiksi — eli samanlaisten tai toisiinsa liittyvien kohteiden ryhmäksi — tunnistamalla samankaltaisia havaintoja tai tiivistämällä dataa pienempään ja helpommin hahmotettavaan muotoon. Tässä alaluvussa tutustutaan ohjaamattoman oppimisen eri tyyppeihin.

### 2.2.1 Klusterointi

Klusterointi on ohjaamattoman oppimisen menetelmä, jonka tavoitteena on ryhmitellä havaintoja siten, että samaan ryhmään kuuluvat havainnot ovat keskenään mahdollisimman samankaltaisia ja eri ryhmiin kuuluvat mahdollisimman erilaisia. Toisin sanoen tavoite on löytää klustereita syötedatasta. Klusterointia käytetään usein hyvin suurien datamäärien analysoimiseen (Mohri ym., 2018).

Tunnetuimpia klusterointimenetelmiä ovat esimerkiksi K-means, joka jakaa havainnot käyttäjän määrittelemään määrään klustereita minimoimalla etäisyyksien neliösumman klusterikeskipisteistä, sekä hierarkkinen klusterointi, jossa havaintoja yhdistetään vähitellen suuremmiksi ryhmiksi niiden samankaltaisuuden perusteella. Lisäksi tiheysperustaiset menetelmät, kuten DBSCAN, tunnistavat klustereita havaintotiheyksien perusteella ja soveltuvat epäsäännöllisen muotoiseen dataan.

### 2.2.2 Dimension vähentäminen

Dimension vähentäminen on tekniikka, jonka tavoitteena on tiivistää monimutkaista tai moniulotteista dataa säilyttäen sen keskeisimmät rakenteet ja informaation. Toisin sanoen pyritään esittämään data pienemmässä ulottuvuudessa siten, että datan keskeinen informaatio ja rakenne on ideaalisti sama kuin alkuperäisessä datassa. Dimension vähentäminen on tärkeää, koska data saattaa olla moniulotteista, jolloin sen käsittely ja ymmärtäminen on erittäin vaikeaa. Jos data on pienemmässä ulottuvuudessa, on sitä myös laskennallisilta resursseilta halvempi ja nopeampi kouluttaa. Koneoppimismallit, jotka on koulutettu suuriulotteisella datalla, generalisoivat – eli yleistävät oppimansa tiedot uusiin, aiemmin näkemättömiin havaintoihin – myös usein huonosti. Dimension vähentäminen voidaan jakaa kahteen tyyppiin: feature selection ja feature extraction. Feature selection on tekniikka, jossa yritämme löytää k-dimensiota jotka antavat eniten informaatiota alkuperäisestä d-dimensiosta. Loppuja (d miinus k) dimensiota ei käytetä. Feature extraction on taas tekniikka, jossa on tavoitteena löytää uusi joukko k-dimensioita, jotka ovat yhdistelmiä alkuperäisistä d-dimensiosta. (Alpaydin, 2020.)

## 2.3 Algoritmit

Koneoppimisessa käytettävät algoritmit muodostavat menetelmien käytännön ytimen: ne määrittävät, kuinka dataa käsitellään, mitä piirteitä siitä opitaan ja minkälaisia päätelmiä voidaan tehdä. Algoritmien valinta riippuu datan luonteesta sekä tutkimuskysymyksistä. Tässä luvussa tutustutaan keskeisiin koneoppimisalgoritmeihin, joita sovelletaan tai voidaan soveltaa saavutettavuustutkimuksessa. Käsitellyt algoritmit ovat joko ohjattuun tai ohjaamattomaan oppimiseen liittyviä algoritmeja. Muut algoritmit ovat jätetty käsittelemättä.

### 2.3.1 K-Nearest Neighbors

K-lähimmän naapurin menetelmä (KNN) on yksinkertainen ei-parametrinen koneoppimisalgoritmi, jota voidaan käyttää luokitteluun ja regressioon. Menetelmä perustuu ajatukseen, että samankaltaisilla havainnoilla on

samankaltaiset vasteet. Kun halutaan ennustaa havaintopisteen arvo, etsitään sen k-lähintä naapuria jonkin etäisyysmitan (esim. euklidinen) perusteella ja luokitellaan piste niiden mukaan.

Luokittelussa uusi havainto saa sen luokan, johon sen k-lähimmät naapurit kuuluvat. Menetelmän etuna on sen yksinkertaisuus, tulkittavuus ja se, että se ei vaadi erillistä mallin parametrisointia tai oletuksia jakaumista. Haittapuolia ovat sen herkkyys k-parametrin valinnalle sekä etäisyysmitan valinnalle.

### 2.3.2 K-Means

K-means on ohjaamattoman oppimisen klusterointialgoritmi, jonka tavoitteena on jakaa havaintoaineisto k-klusteriin siten, että saman klusterin eri havainnot ovat keskenään mahdollisimman samankaltaisia ja eri klustereiden havainnot mahdollisimman erilaisia. Algoritmi minimoi havaintojen etäisyyksien neliösumman klusterikeskipisteistä.

Algoritmi toimii iteratiivisesti: aluksi valitaan k alkuperäistä klusteripistettä (satunnaisesti tai jollakin muulla tavalla). Tämän jälkeen jokainen havainto liitetään lähimpään keskipisteeseen. Kun kaikki havainnot on liitetty, klusterin keskipisteet lasketaan uudelleen klusterin omien havaintojen keskiarvona. Tämä prosessi toistetaan, kunnes keskipisteet eivät enää muutu tai jokin pysäytyskriteeri toteutuu.

K-means etuna on sen yksinkertaisuus ja tehokkuus suurillakin aineistoilla. Haasteita ovat tarve määrittää klustereiden määrä (k) etukäteen sekä herkkyys alkuarvoille.

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to $k$ random $\boldsymbol{x}^t$
Repeat
    For all $\boldsymbol{x}^t \in \mathcal{X}$
        $$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$
    For all $\boldsymbol{m}_i, i = 1, \ldots, k$
        $$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$
Until $\boldsymbol{m}_i$ converge

Kuva 1: k-means pseudokoodina (Alpaydin, 2020).

### 2.3.3 Päätöspuu ja satunnaismetsä

Päätöspuut ovat koneoppimismalleja, jotka esittävät päätöksenteon puumaisena rakenteena. Jokainen solmu (engl. node) vastaa ehtoa jollekin

selitettävälle muuttujalle. Oksat (engl. branch) edustavat näiden ehtojen lopputuloksia ja lehtisolmut sisältävät ennusteen, joka on luokka tai arvo. Päätöspuut ovat yksinkertaisia ja helposti tulkittavia, mutta yksittäisinä malleina ne ovat alttiita ylioppimiselle (engl. overfitting) ja datan satunnaiselle vaihtelulle.

Satunnaismetsä (engl. Random Forest, RF) on kokonaisuusmenetelmä (engl. ensemble method), jossa rakennetaan suuri määrä päätöspuita ja yhdistetään niiden ennusteet. Jokainen puu opetetaan hieman erilaisella otoksella datasta (bootstrap-näyte), ja jokaisessa puun jakokohdassa käytettävissä on vain satunnainen alajoukko muuttujia, jolloin yksittäisten vahvojen piirteiden hallitseva vaikutus pienenee ja puiden rakenne monipuolistuu. Lopullinen tulos saadaan tyypillisesti äänestyksellä tai puiden ennusteiden keskiarvona. RF menetelmän etu ovat hyvä ennustustarkkuus, kyky käsitellä suuria ja moniulotteisia aineistoja sekä vankkuus poikkeaville havainnoille.

### 2.3.4 Neuroverkko

Keinotekoiset neuroverkot ovat löyhästi aivojen rakenteeseen perustuva koneoppimismalli. Eräs näihin malleihin kuuluva malli on monikerroksinen perceptron (engl. Multilayer perceptron, MLP). Nämä mallit ovat syväoppimisen perusta ja niitä voi käyttää niin regressio- kuin luokittelutehtävissäkin.

MLP koostuu syötekerroksesta, piilotetuista kerroksista ja tuloskerroksesta. Syötekerroksessa jokainen neuroni vastaa yhtä syötemuuttujaa. Piilotettuja kerroksia voi olla mielivaltainen määrä ja ne prosessoivat syötekerroksen informaatiota. Tuloskerros tulostaa lopullisen luokan tai tuloksen.

MLP toimii siten, että syöte välittyy kerroksittain verkon läpi kohti tuloskerrosta. Jokainen neuroni laskee ensin saamansa syötteen painotetun summan ja lisää siihen bias-termin. Tämän jälkeen tulos muunnetaan aktivaatiofunktion avulla, mikä tuo malliin epälinearisuutta ja mahdollistaa monimutkaisten riippuvuuksien oppimisen. Kun syöte liikkuu kerroksesta toiseen, verkko rakentaa yhä abstraktimpia esityksiä syötteestä.

Koulutuksessa verkko oppii säätämällä sisäisiä parametrejaan eli painojaan. Ensin tehdää niin kutsuttu forward propagation, jossa verkko tuottaa arvion halutusta tuloksesta. Tätä arviota verrataan opetusdataan valitun hukkafunktion, kuten MSE:n, avulla. Tämän jälkeen virhe kuljetetaan taaksepäin verkossa niin kutsutulla backpropagation-algoritmilla, joka laskee jokaisen painon vaikutuksen virheeseen. Lopuksi optimoija, kuten Adam, päivittää painoja pienentämään virhettä. Tätä prosessia toistamalla verkko oppii kuvaamaan syötteiden ja tavoitemuuttujien välistä suhdetta.

# 3 Saavutettavuustutkimuksen teoreettinen tausta

Saavutettavuus on keskeinen käsite liikenne- ja aluesuunnittelun sekä maantieteen tutkimuksessa. Saavutettavuustutkimus on tutkimusta, jossa pyritään selvittämään kuinka helposti ihmiset, palvelut ja toiminnot ovat saavutettavissa eri sijainneissa. Hansen (1959) klassisesti määritteli saavutettavuuden henkilön mahdollisuutena osallistua toimintoihin tietyssä sijainnissa. Tämän jälkeen saavutettavuuden käsite on laajentunut ja sitä on sovellettu niin liikennejärjestelmien arviointiin, yhdyskuntarakenteen suunnitteluun, kuin sosiaalisen tasa-arvon tutkimukseen.

Saavutettavuuden käsite on monitulkintainen ja sillä onkin eri määritelmiä riippuen mistä näkökulmasta sitä tarkastellaan. Handy ja Niemeier (1997) huomauttavatkin, että ei ole olemassa yhtä parasta tapaa lähestyä saavutettavuutta, koska eri tilanteet ja käyttötarkoitukset vaativat erilaisia lähestymistapoja. Hansenin määritelmän lisäksi Dalvi ja Martin (1976) määrittelivät saavutettavuuden, siten, kuinka helposti mihin tahansa maankäyttöön liittyvään toimintaan voidaan päästä käyttämällä tiettyä liikennejärjestelmää. Burnsin (1979) mukaan saavutettavuus määritellään yksilön vapautena päättää osallistumisestaan eri aktiviteetteihin. Nämä erilaiset määritelmät osoittavat, että saavutettavuus ei ole vain fyysinen tai spatiaalinen ilmiö, vaan myös sosiaalinen.

## 3.1 Saavutettavuuden komponentit

Saavutettavuus muodostuu useista toisiinsa sidoksissa olevista osatekijöistä, jotka kuvaavat sekä toimintojen sijaintia, että niiden saavutettavuutta erilaisilla tavoilla. Geursin ja Van Ween (2004) mukaaan saavutettavuus voidaan jakaa neljään pääkomponenttiin: maankäyttö-, liikenne-, aika- ja yksilökomponenttiin. Nämä komponentit muodostavat perustan lähes kaikille saavutettavuusmittareille.

Geursin ja Van Ween (2004) mukaan maankäyttökomponentti kuvaa maankäyttösysteemiä, johon kuuluu (a) kunkin kohteen tarjoamien mahdollisuuksien määrä, laatu ja spatiaalinen jakautuminen (esimerkiksi työpaikat, kaupat ja terveyspalvelut), (b) näiden palveluiden kysyntä lähtöalueilla eli alueella missä asukas asuu ja (c) kysynnän ja tarjonnan kohtaaminen eri mahdollisuuksille, mikä voi johtaa kilpailuun rajallisen kapasiteetin toiminnoista, kuten työ- ja sairaalapaikoista.

Toinen Geursin ja Van Ween (2004) esittämä komponetti on liikennekomponentti, joka kuvaa liikennejärjestelmää, joka ilmenee yksilölle koituvana

haittana matkan tekemisestä lähtöpaikan ja kohteen välillä tiettyä kulkutapaa käyttäen. Tähän haittaan sisältyy matkaan kuluva aika, kustannukset ja vaiva. Tämä haitta syntyy kysynnän ja tarjonnan vuorovaikutuksesta. Infrastruktuurin tarjonta käsittää sen sijainnin ja ominaisuudet ja kysyntä puolestaan liittyy sekä henkilö- että tavaraliikenteeseen.

Geursin ja Van Ween (2004) mukaan aikakomponentti kuvaa ajallisia rajoitteita, eli esimerkiksi mahdollisuuksien saatavuutta päivän eri aikoina ja aikaa, joka yksilöillä on käyttää eri aktiviteetteihin.

Geurisin ja Van Ween (2004) viimeinen komponentti on yksilökomponentti, joka kuvaa (a) yksilöiden tarpeita, jotka riippuvat yksilön iästä ja koulutuksesta ym., (b) kykyjä, jotka riippuvat henkilön fyysisestä kunnosta ja käytössä olevista kulkutavoista ym. ja (c) mahdollisuuksia, jotka riippuvat henkilön tuloista ja matkabudjetista ym. Nämä ominaisuudet vaikuttavat henkilöiden pääsyyn eri kulkutapoihin ja spatiaalisesti jakautuneisiin mahdollisuuksiin. Nämä ominaisuudet myös voivat vaikuttaa saavutettavuuden kokonaisarvoon.

## 3.2 Saavutettavuuden mittarit

Saavutettavuuden mittarit mittaavat saavutettavuutta. Geursin ja Van Ween (2004) mukaan mittarin tulisi ottaa huomioon kaikki saavutettavuuden komponentit, mutta käytännössä mittarit ottavat huomioon yhden tai useamman riippuen siitä, mistä näkökulmasta saavutettavuutta tarkastellaan. Saavutettavuusmittarit voidaan jakaa karkeasti neljään pääluokkaan: infrastruktuuriin perustuvaan, sijaintiin perustuvaan, ihmisiin perustuvaan ja hyötyyn perustuvaan mittariin.

Geursin ja van Ween (2004) mukaan infrastruktuuripohjaiset mittarit kuvaavat liikenneinfrastruktuurin – havaittua tai simuloitua – suorituskykyä tai palvelutasoa, kuten ruuhkatasoa. Tämän tyyppisiä mittareita käytetään usein liikennesuunnittelussa. Näihin mittareihin ei sisälly maankäytön komponentti.

Sijaintipohjaiset mittarit puolestaan tarjoavat makrotason analyysin saavutettavuudesta. Ne analysoivat saavutettavuutta tietyssä sijainnissa ja kuvaavat tason, jolla sijainti on saavutettava alueelleisesti jakautuneisiin toimintoihin. Nämä mittarit voidaan jakaa alaluokkiin: (1) Etäisyys- ja ääriviivamittareihin, jotka määrittelevät missä määrin paikat ovat yhteydessä toisiinsa ja mahdollisuuksien määrän, joihin päästään annetun matka-ajan, etäisyyden tai kustannusten puitteissa, (2) potentiaali/painovoima mittareihin, jotka arvioivat mahdollisuuksien saavutettavuutta joltain vyöhykkeeltä toisiin vyöhykkeisiin. Tällaisia mittareita on muun muassa HAM.

Henkilöpohjaiset mittarit analysoivat saavutettavuutta yksilötasolla. Mittarit perustuvat Hägerstrandin (1970) aikamaantieteeseen (engl. space-time geography), joka tarkastelee yksilön toimintavapauden rajoituksia ympäristössä, kuten pakollisten toimintojen sijaintia ja kestoa.

Viimeisenä Geurs ja Van Ween (2004) esittelevät hyötyyn (engl. utility) perustuvat mittarit. Nämä mittarit analysoivat taloudellista hyötyä, jonka ihmiset saavat pääsystä alueellisesti jakautuneisiin toimintoihin.

## 3.3 Spatiaalinen interaktiomalli

Spatiaalinen vuorovaikutusmalli (engl. spatial interaction model, SIM) kuvaa, kuinka vuorovaikutus kahden paikan välillä riippuu niiden ominaisuuksista ja niiden välisistä etäisyyksistä tai matkakustannuksista. Salzen ym., (2011) mukaan spatiaaliset vuorovaikutusmallit pohjautuvat sosiaaliseen fysiikkaan (engl. social physics) ja olettavat vastaavuuksia fysikaalisten (esim. Newtonin gravitaatiolain) ja sosiaalisten ilmiöiden (esim. muuttoliikkeen) välillä. Wilsonin (1967, 1970) mukaan SIM on entropiamaksimoinnin tasapainoratkaisu (engl. equilibrium solution). Fotherigamin ja O'Kellyn (1989) mukaan SIM:it ennustavat vuorovaikutuksen hyvin, kun matriisi, joka sisältää lähtöpaikat ja kohdepaikat, on tiedossa. Tällaisia malleja kutsutaan doubly-constrained -malleiksi.

Doubly-constrained malli määritellään:

$$T_{ij} = A_i O_i B_j D_j f(t_{ij}) \tag{1}$$

missä $T_{ij}$ on virrat eli matkat lähtöpaikasta i kohteeseen j, $O_i$ on ulosvirtaukset eli matkat lähtökohteesta i, $D_j$ on sisäänvirtaukset eli matkat, jotka vetäytyvät kohteeseen j, $f(t_{ij})$ on impendanssifunktio ja $A_i$ ja $B_j$ ovat niin sanottuja tasapainottavia tekijöitä.

Wilsonin mukaan tasapainottavalla tekijällä $A_i$, joka esiintyy production-constrained SIM:issä:

$$T_{ij} = A_i O_i D_j f(t_{ij}) \tag{2}$$

On ominaisuus:

$$A_i = \frac{1}{\sum D_j f(t_{ij})} \tag{3}$$

$A_i$ on siis käytännössä saavutettavuuden käänteisluku (engl. inverse). Hansenin (1972) mukaan Hansenin saaavutettavuusmuotoilu voidaan linkittää systeemin entropiakuvioihin (engl. entropy pattern), koska SIM eq1 nousee entropian maksimoinnin tasapainoratkaisusta.

## 3.4 Hansenin saavutettavuusmalli

Yksi tunnetuimmista saavutettavuuden mittareista on Hansenin saavutettavuusmalli (engl. Hanses Accessibility Model, HAM). Hansenin saavutettavuusmalli on potentiaalimittari. Hansenin (1959) mukaan potentiaalinen saavutettavuus $ACC_i$ alueesta i kuvaavat alueen j mahdollisuuksia kaavalla:

$$ACC_i = \sum_j W_j t_{ij}^{-\gamma} \tag{4}$$

Jossa $ACC_i$ on alueen i saavutettavuus, joka on summa alueen mahdollisuuksista $W_j$, kuten ostosmahdollisuudet, kerrottuna aikahajoamisfunktiolla $t_{ij}^{-\gamma}$. Hansenin (1959) mukaan aikahajoamisfunktio ottaa huomioon alueen houkuttelevuuden negatiivisen potenssimuodon mukaan. Handy ja Niemeierin (1997) mukaan negatiivinen exponenttifunktio on eniten käytetty impendanssifunktio ja se kytkeytyy myös parhaiten matkustuskäyttäytymisteoriaan (engl. travel behaviour theory). Kuitenkin myös muita impendanssifunktioita voi käyttää kuten, logistista tai Gaussin funktioa.

# 4 Koneoppiminen saavutettavuustutkimuksessa

Koneoppimismenetelmien käyttö kaupunki- ja saavutettavuustutkimuksessa on lisääntynyt viime vuosina, kun saatavilla olevien paikkatietoaineistojen ja liikennedatan määrä on kasvanut merkittävästi.

Koneoppimismenetelmät tarjoavat vaihtoehdon perinteisille saavutettavuusmalleille. Näiden mallien avulla saavutettavuutta voidaan mallintaa tehokkaasti eri riippuvuuksista ja löytää datasta erilaisia rakenteita, joita perinteisillä menetelmillä ei pystytä välttämättä kuvaamaan. Casalin ym. (2022) mukaan useat tutkimukset osoittavatkin, että esimerkiksi tukivektorikoneet (SVM), satunnaismetsät (RF) ja boostatut päätöspuut (DT) ovat osoittautuneet erityisen tehokkaiksi kaukokartoitusdatan luokittelussa.

Tässä luvussa tutustutaan siihen, minkälaisia koneoppimismenetelmiä on käytetty saavutettavuustutkimuksessa, minkälaisia tutkimuskysymyksiä niillä on pyritty ratkaisemaan ja minkälaisia tuloksia, etuja ja haasteita näihin menetelmiin liittyy.

## 4.1 Koneoppimismenetelmien käyttökohteet saavutettavuudessa

Koneoppimismenetelmiä on viime vuosina sovellettu monipuolisesti saavutettavuustutkimuksessa, vaikka ala onkin vielä hyvin nuori. Tässä alaluvussa tutustumme näihin tutkimuksiin ja niiden konteksteihin. Tutkimusten tavoitteena on ollut kehittää uusia menetelmiä, jotka kykenevät käsittelemään suuria ja monimuotoisia paikkatietoaineistoja ja tunnistamaan saavutettavuuteen vaikuttavia piileviä rakenteita ja alueellisia eroja.

### 4.1.1 Koneoppimisen soveltaminen palveluiden saavutettavuuteen

Khosravi Kazazi ym. (2022) kehittivät koneoppimisen hybridimallin, jonka tarkoituksena oli kartoittaa terveyspalveluiden saavutettavuutta Iranin Isfahanin kaupungissa. Tutkimuksessa yhdistettiin ohjaamaton ja ohjattu oppiminen. Ohjaamattomat ML-metodit olivat klusterointimenetelmiä: K-Means klusterointi, Agglomerative K- Means klusterointi ja Bisecting K-Means klusterointi.

Ensimmäisessä tutkimusvaiheessa alueen jokaiselle blokille laskettiin etäisyys terveyspalveluihin. Etäisyydet määritettiin OpenStreetMapin

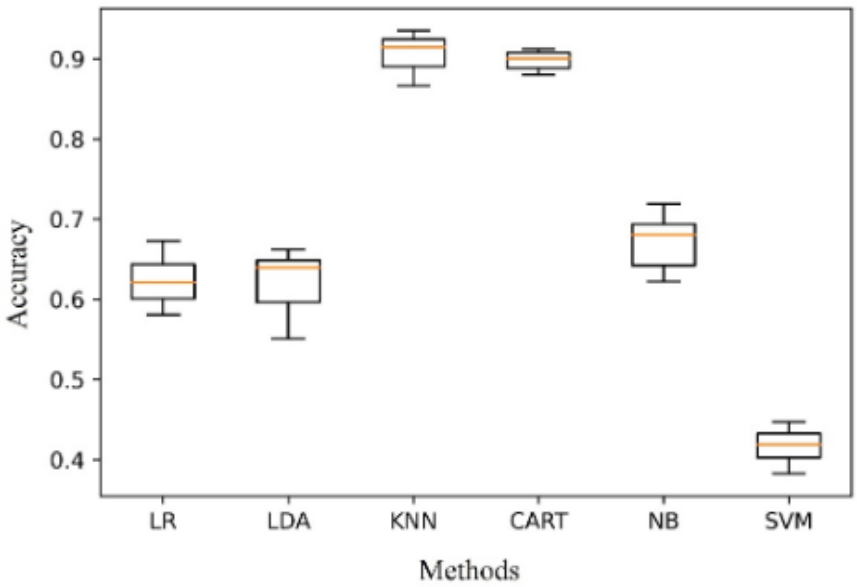
katuverkon pohjalta käyttäen lyhimmän polun algoritmia (Dijkstra), jossa saavutettavuus perustui todellisiin kulkureitteihin, ei linnuntiehen.

Tämän jälkeen saavutettavuusmuuttujat normalisoitiin min—max – skaalauksella, jotta kaikki etäisyydet olisivat samalla vaihteluvälillä, eikä mikään yksittäinen muuttuja hallitsisi tulosta. Sitten muuttujille tehtiin Yeo—Johnson-power transformation-muunnos, jonka tarkoituksena oli korjata jakaumien vinoutta ja parantaa klusterointialgoritmien toimintaa, sillä monet menetelmät toimivat luotettavammin, kun data muistuttaa normaalijakaumaa. Lopuksi muuttujille laskettiin Shannonin entropiaan perustuvat painot, jotta analyysi huomioisi eri muuttujien informatiivisuuden – suuremman hajonnan omaavat muuttujat erottelevat alueita paremmin ja saavat siksi suuremman painon saavutettavuusindeksin laskennassa.

Vielä ennen klusterointia klustereiden määrä määritettiin kyynärpäämenetelmällä (engl. elbow method). Menetelmä perustuu niin sanottuun within-cluster sum of squares -arvoon (WCSS), joka kuvaa kuinka lähellä samaan klusteriin kuuluvat havaintopisteet ovat oman klusterinsa keskipistettä. Kun klustereiden määrää lisätään, WCSS-arvo pienenee, koska data voidaan jakaa yhä pienempiin homogeenisiin ryhmiin. Pieneneminen ei kuitenkaan jatku lineaarisesti, vaan jossain vaiheessa lisäklustereiden tuoma arvo muuttuu vähäiseksi. Tätä kohtaa kutsutaan kyynärpääksi, ja se on arvo, joka valitaan klustereiden määräksi. Tässä tutkimuksessa kyynärpää oli kohdassa k=5.

Tutkimusalue jaettiin viiteen klusteriin saavutettavuuden perusteella. Nämä klusterit siis edustivat saavutettavuuden eri tasoja: erittäin hyvä, hyvä, keskitasoinen, heikko ja erittäin heikko. Kunkin klusterointimenetelmän avulla muodostettiin oma saavutettavuuskartta, ja tuloksia verrattiin sen perusteella, kuinka selvästi menetelmät erottelivat saavutettavuusluokat. Tämän jälkeen klusteroinnin tuloksia käytettiin opetusdatan pohjana ohjatulle koneoppimiselle, jotta voitiin luokitella saavutettavuusluokat tutkimusalueella automaattisesti. Klusteroinnista saatu data jaettiin koulutusdataan (80 %) ja testidataan (20 %). Kuutta eri ohjatun oppimisen algoritmia kokeiltiin: lineaarinen regressio (LR), lineaarinen diskriminanttianalyysi (LDA), k-nearest neighbours (KNN), CART, naive Bayes (NB) ja tukivektorikone (SVM).

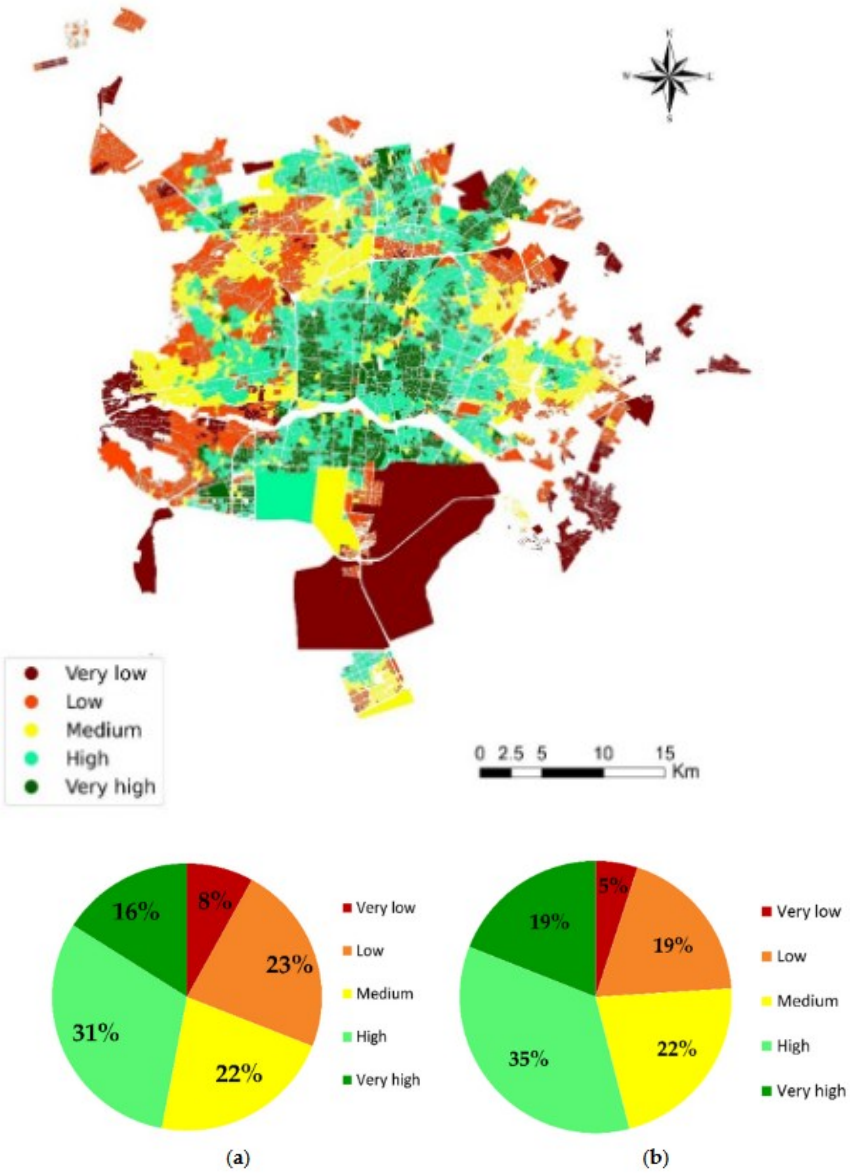
Tämän jälkeen näiden tuloksia vertailtiin. Lopuksi KNN ja CART tuottivat parhaat tulokset.




Kuva 2: Eri algoritmien tarkkuus (Khosravi Kazazi ym., 2022).

KNN valittiin lopulliseksi malliksi CART:in sijaan, sillä se on yksinkertaisempi algoritmi ja se on ei-parametrinen — eli se ei oleta datan olevan mistään tietystä jakaumasta.

Lopuksi KNN-ML-metodilla luotiin saavutettavuuskartta ja laskettiin jokaisen luokan suhteellinen osuus.



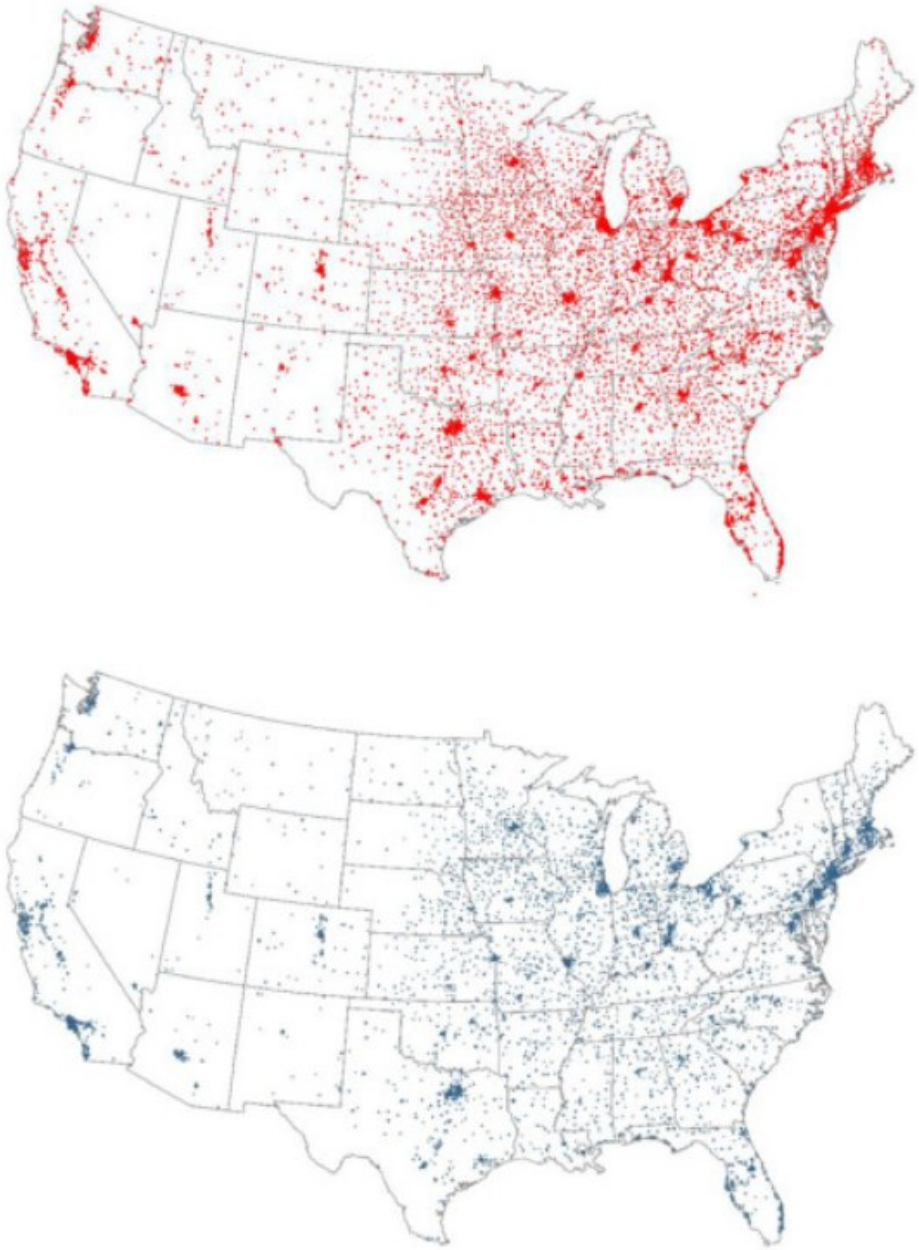

Kuva 3: Ennustetut luokat ja niiden suhteelliset osuudet (Khosravi Kazazi ym., 2022).

Tutkumus osoittaa, että ohjaamattoman oppimisen menetelmät toimivat hyvin palveluiden saavutettavuuden luokittelussa silloin, kun valmiita luokkia ei ole määritelty. KNN ja CART antoivat parhaat tulokset tarkkuudessa (engl. accuracy). Tutkimusalueella 47 % alueista saavutettavuus on hyvää, 22 % keskinkertaista ja 31 % huonoa. Kartasta nähtynä keskustassa saavutettavuus oli hyvää tai erittäin hyvää ja reuna-alueilla heikompaa.

Toisessa palveluiden saavutettavuuteen liittyvässä tutkimuksessa mitattiin ja kartoitettiin Yhdysvaltojen mannearalueen korkealaatuisten (4–5 tähteä) hoivakotien saavutettavuutta kaikille maakunnille ja analysoitiin, mitkä väestö- ja aluerakenteiden tekijät selittävät saavutettavuuden vaihtelua.

Tutkimuksen tavoitteena oli siis 1) mitata ja kartoittaa Yhdystaltojen maakuntien kokonaan kattava spatiaalinen saavutettavuus hyvin luokiteltuihin hoivakohteihin ja 2) löytää olennaisimmat sosiodemogragfiset muuttujat, jotka vaikuttavat saavutettavuustasoihin.

Datana käytettiin 1) kaikkia saatavilla olevia Yhdysvaltojen hoivakotien tähtiluokituksia, jotka jaoteltiin hyviin ja huonoihin ja 2) niiden koordinaatteja, jotka geokoodattiin lat—long muotoon. Sen jälkeen maakunnat, joissa hoivakoti sijaitsi, liitettiin dataan. Hoivakoteja oli yhteensä 15 215 joista hyviä (4–5 tähteä) oli 6904 (45 %).



Kuva 4: Kaikki hoivakodit (punainen) ja hyvät hoivakodit (sininen) Yhdysvalloissa (Reddy ym., 2022).

Toinen keskeinen aineistolähde oli American Community Survey (ACS) vuosilta 2012–2016, joka tarjoaa maakuntatason arviot sosiodemografisesta datasta. Datassa oli mukana muuttujat: ”comparative economic characteristics”, ”ACS demographic and housing estimates”, “age and sex” ja “Household and families”. Joitakin lisämuuttujia lisättiin, kuten maakuntien asukastiheys.

Saavutettavuusmitta hyvään hoivakotiin laskettiin gravitaatiopotentiaalimallilla. Jokaiselle maakunnalle laskettiin populaatiopainotettu systeeminlaajuinen saavutettavuus (engl. population-based system-wide spatial accessibility) hyvän arvion saaneisiin hoivakoteihin. Tämä laskettiin kaavalla:

$$SA_i = \sum_j \left[ \left( \frac{n_j}{p_i} \right) \left( \frac{1}{max(d_{ij}^2, 1)} \right) \right] \tag{5}$$

Jossa $n_j$ kuvaa vuoteiden lukumäärää jokaisessa hyväksi luokitellussa hoivakodissa j, $p_i$ kuvaa yli 65-vuotiaiden henkilöiden lukumäärää maakunnassa i ja $d_{ij}$ kuvaa etäisyyttä kilometreinä maakunnan i painopisteen (engl. centroid) ja hoivakodin j välillä. Mitä suurempi SA:n arvoksi saadaan, sitä parempi saavutettavuus maakunnalla on hyvään hoivakotiin.

Seuraava askel analyysissä oli luoda selittävä malli (engl. expalnatory model), jonka tarkoituksena oli löytää yhteyksiä eriarvoisuutta kuvaavien mittareiden ja hyvään hoivakotiin pääsyn välillä. Tämä tapahtui poistamalla ensin datasta yli 65-vuotiaat, koska tämä oli osa saavutettavuusmittarin laskentaa. Joissain ACS taulukoissa esiintyneet virhemarginaalit (engl. margin of error) poistettiin ja malliin otettiin mukaan vain piste-estimaatit. Kaksoiskappaleina esiintyvät muuttujat yhdistettiin ja äärimmäisen korreloivat (~0.99) muuttujat poistettiin. Lopullisesssa aineistossa oli 472 muuttujaa ja 3074 maakuntaa. Koneoppimismenetelmät saivat työkseen valita sopivat muuttujat näistä.

Tavoitteena oli kehittää malli, joka tunnistaa kaikkein tärkeimmät sosiodemografiset tekijät, jotka liittyvät maakuntatason saavutettavuuteen. Mallinnus aloitettiin kokeilemalla perinteistä pienimmän neliösumman menetelmää (OLS), mutta se osoittautui riittämättömäksi, koska muuttujia oli hyvin paljon (472). Monet muuttujista koerreloivat vahvasti keskenään ja OLS ei kykene tällöin tekemään automaattista muuttujavalintaa. Lisäksi havaittiin, että OLS-malli ei huomioinut eräitä keskeisiä muuttujia, jotka myöhemmin osoittautuivat tärkeiksi.

Tämän vuoksi käytettiin satunnaismetsä (RF) -menetelmää arvioimaan, mitkä muuttujat näyttivät olevan merkityksellisiä ilman ennakkorajoitteita. RF paljasti, että erityisesti Amerikan alkuperäiskansoihin liittyvät väestönosuudet selittivät saavutettavuuden vaihtelua vahvasti. Koska nämä muuttujat eivät olleet mukana alkuperäisessä OLS-mallissa, kävi selväksi, että pelkkä perinteisiin tilastomenetelmiin perustuva muuttujavalinta ei

riittänyt. Tarvittiin menetelmä, joka antaisi datan itse määrittää olennaiset tekijät.

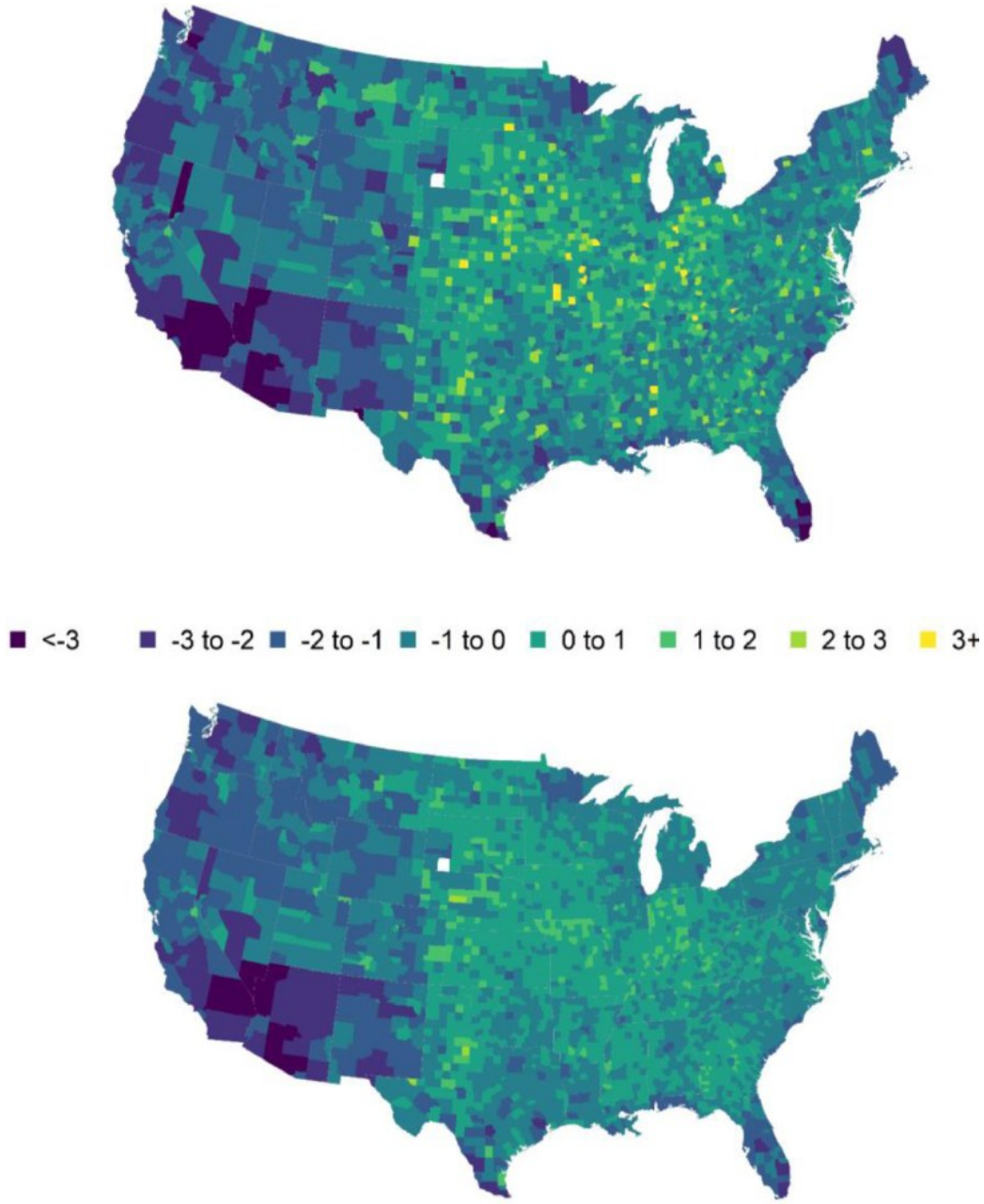
Tämän saavuttamiseksi siirryttiin käyttämään koneoppimismenetelmää, joka pystyy käsittelemään suuren määrän muuttujia ja tekemään systemaattisen muuttujavalinnan rangaistustermien avulla. Ensin puuttuvat tiedot täydennettiin RF-imputoinnilla, koska ACS-aineistossa oli ei-satunnaista puuttuvaa dataa erityisesti Kalliovuorten alueella. RF-imputointi soveltuu hyvin tällaisiin tilanteisiin, koska se hyödyntää monimutkaisia muuttujien välisiä riippuvuuksia ja tuottaa keskimäärin tarkemmat imputoidut arvot.

Imputoinnin jälkeen käytettiin LASSO-regressiota (Least Absolute Shrinkage and Selection Operator) valitsemaan aineiston kaikista 472 muuttujasta vain ne tekijät, jotka selittivät saavutettavuuden vaihtelua tilastollisesti merkittävästi. LASSO lisää malliin L1-rangaistustermin, joka pakottaa epäolennaisten muuttujien kertoimet nollaan ja tuottaa siten parsimonisen eli niukkamuutujaisen ja tulkittavan mallin. Yhdistettynä tätä menetelmää kutsuttiin nimellä imputoitu LASSO, ja se mahdollisti sekä puuttuvan datan käsittelyn että tehokkaan muuttujavalinnan.

LASSO on rangaistu regressiomenetelmä, joka Reddyn ym. (2022) mukaan parantaa mallin tulkittavuutta ja ennustettavuutta poistamalla mallista ne muuttujat, jotka eivät selitä ilmiötä merkittävästi. Imputoidun LASSO lähestymistavan todetaan täten tuottavan paremman mallin kuin pelkän satunnaismetsän käytön ja vähentävän puuttuvan datan aiheuttamaa vinoumaa erityisesti, kun datan puuttuminen ei ole satunnaista, kuten tässä aineistossa, jossa Kalliovuorten tiedot olivat puutteellisia.

Kaikki muuttujat keskitettiin ja skaalattiin keskiarvon ja keskihajonnan suhteen, joka tekee näistä z-arvoja. Aineisto jaettiin koulutus- ja testijoukkoon, siten että 75 % datasta oli koulutusdataa ja 25 % testidataa.

Analyysi toteutettiin Post-LASSO -menetelmällä. Post-LASSO korjaa LASSON aiheuttamaa estimaattien vinoumaa ja mahdollistaa luottamusvälien (engl. confidence interval) laskemisen muuttujavalinnan jälkeen. Tämä tekee tuloksista hyvin tulkittavia ja luotettavia. Reddyn ym. mukaan tämä on esimmäinen kerta, kun Imputed Post-LASSO -menetelmää on käytetty.




Kuva 5: Maakuntien laskettu saavutettavuus (ylhäällä) ja luodun Ml-menetelmän antamat ennusteet (alhaalla). Mitä suurempi arvo, sitä parempi saavutettavuus. (Reddy ym., 2022.)

Paras saavutettavuus havaittiin keskilännessä ja Kalliovurten itäosissa. Heikoin saavutettavuus taas Tyynenmeren rannikko-osavaltioissa sekä Yhdysvaltojen lounaisosissa.

Imputed Post-LASSO loi onnistuneesti niukkalukuisen (parsimonisen) mallin, joka vähensi muuttujien määrää reilusti. 28 muuttujaa 472 muuttujasta valittiin. Mallin selitysaste ($R^2$) oli 0.613 testidatalla, joka viittaa suhteellisen

hyvään tarkkuuteen. Tutkimus osoitti, että saavutettavuus vaihtelee alueellisesti ja on yhteydessä väestöntiheyteen, väestön etniseen taustaan ja sosioekonomisiin tekijöihin. Koneoppimismenetelmä (imputoitu post-lasso) osoittautui tehokkaaksi tavaksi tunnistaa saavutettavuuden kannalta merkittävimmät muuttujat.

## 4.2 Koneoppimisen yhdistäminen saavutettavuusteorioihin

Saavutettavuuden ymmärtäminen on keskeistä kaupunkisuunnittelussa ja aluekehityksessä, koska se kuvaa, kuinka helposti ihmiset voivat saavuttaa erilaisia palveluita, työpaikkoja ja muita kohteita. Hadjidimitriou ym. (2025) mukaan saavutettavuus voidaan nähdä piilevänä ajorattaana talouskasvulle ja kehitykselle. Saavutettavuus vaikuttaa suoraan yksilön arjen sujuvuuteen ja alueiden taloudelliseen ja sosiaaliseen elinvoimaan. Perinteiset saavutettavuusmallit, kuten HAM, tarjoavat vakiintuneen tavan mitata saavutettavuutta, mutta nykyaikaiset datalähteet ja laskennalliset menetelmät, kuten koneoppiminen, avaavat uusia mahdollisuuksia saavutettavuuden analysoimiseen ja ennustamiseen.

Hansenin saavutettavuusmallin ja koneoppimisen yhdistäminen kaupunkien saavutettavuuden analysoimiseksi on potentiaalinen korvike perinteisille menetelmille ja se voisi mahdollistaa saavutettavuuden analysoinnin silloinkin, kun virtausdata (engl. mobility flow data) on puutteellista. Hadjidimitriou ym. tutkivat, voivatko ML- mallit, kuten RF ja NN toistaa HAM-mallin käyttäytymisen ja siten ennustaa saavutettavuutta erityisesti silloin kun virtausdataa ei ole saatavilla.

Empiirisenä alustana käytettiin kahta Italian pohjoisosan aluetta, Lombardiaa ja Emilia Romagnaa, analysoiden alueiden sisäisten kaupunkien välistä saavutettavuutta vuosina 2011 ja 2020. Tavoite oli mallintaa vuoden 2011 saavutettavuutta ja sen jälkeen ennustaa vuoden 2020 saavutettavuus kun virtausdata puuttuu Covid-19 pandemian vuoksi. Kahteen ongelmaan etsittiin vastausta luomalla HAM-ML hybridi malli: 1) voiko ML-mallit oppia datan perusteella HAM:in teoreettisen perustan ja 2) voiko ML tarjota ennennäkemättömän ja uuden näkökulman HAM:ista hahmotunnistuksen (engl. Pattern recognition) avulla. HAM-ML-mallin luomisessa ensimmäinen vaihe oli luoda HAM kummastakin kaupungista vuonna 2011 spatiaalisen interaktiomallin (SIM) kalibroinnin avulla. Tämän jälkeen luotiin ML-malli, joka ennustaa saavutettavuuden vuonna 2020.

Tavoitteeseen pääsyyn eli arvottamaan (engl. rank) kunnat niiden HAM:in mukaan, ML-algoritmi tulee kouluttaa HAM-yhtälöllä, jossa käytetään sisäänvirtausta $D_j$, matka-aikaa $t_{ij}$ ja kalibroitua γ-arvoa eli:

$$ACC_i = \sum_j D_j t_{ij}^{-\gamma} \tag{6}$$

γ -arvo saadaan kalibroimalla SIM-yhtälö 1 vuoden 2011 käyttäen sisäänvirtauksia ja matka-aikaa. Tämä arvo sitten lisätään HAM-yhtälöön 6. Tämä kalibroitu HAM, joka sisältää sisäänvirtaukset $D_j$ ja kalibroidun γ-arvon on kohde, jota ML-malli yrittää replikoida. Kalibroiduiksi γ-arvoiksi saatiin Lombardiaan 1,640 ja Emilia Romagnaan 1,455. Nämä γ-arvot syötetään HAM-yhtälöön, joka on maalifunktio ML-mallille.

Taulukko 1: kalibroidut γ-arvot

| **Italian alue** | **γ** |
|---|---|
| Lombardia | 1,640 |
| Emilia Romagna | 1,455 |

ML-malleiksi valittiin NN ja RF. Nämä kumpikin koulutettiin, jotta voidaan löytää, kumpi malli matkii paremmin Hansenin saavutettavuusrankingia. Mallit koulutettiin kymmenen kertaa ja niistä valittiin paras. Etenkin neuroverkon koulutuksessa, koulutuksen tulee tapahtua monta kertaa, koska lopputulos, johon päästään on aina lokaali optimi, ei globaali. Koulutuksessa maalimuuttuja on aiemmin kehitetty HAM. RF käytti 1000:tta estimaattoria, eli tulos on tuhannen DT:n keskiarvo. NN oli feedforward -verkko (tai multilayer perceptron, MLP) ja siinä on yksi piilotettu kerros (engl. hidden layer), jossa on 800 neuronia. Piilotetun kerroksen aktivointifunktiona käytettiin ReLU:a (rectified linear unit) ja tuloskerroksessa käytettiin lineaarista aktivaatiota. ReLU määritellään:

$$ReLU(x) = \max(0, x) \tag{7}$$

Koulutusalgoritmiksi valittiin Adaptive Moment Estimation (adam) ja hukkafunktioksi (engl. loss function) MSE. RF ja NN jakavat samat syötemuuttujat. Ensimmäinen muuttuja on työntekijöiden lukumäärä ($E_i$) kunnassa i. Toinen on HAM eq.4 jossa $w_j = E_j$ ja γ = 1 eli:

$$Att_i = \sum_{j, i \neq j} \frac{E_j}{t_{ij}} \tag{8}$$

Jossa $Att_i$ mittaa kohteiden houkuttelevuutta HAM eq.4 mukaan. HAM (eq6), joka aiemmin kehitettiin vuodelle 2011, on maali, jota ML-malli, joka rakennetaan, yrittää replikoida. Tähän malliin, syötetään $E_i$ ja eq.8 jolloin se oppii eq.6:n saavutettavuuden ja γ-arvon, toisin sanoen HAM-ML koulutus vaiheessa on:

$$Acc_{2011} = ML(E_{2011}, Att_{2011}) \tag{9}$$

Kun malli on koulutettu, voidaan ennustaa 2020 saavutettavuusarvoja seuraavasti:

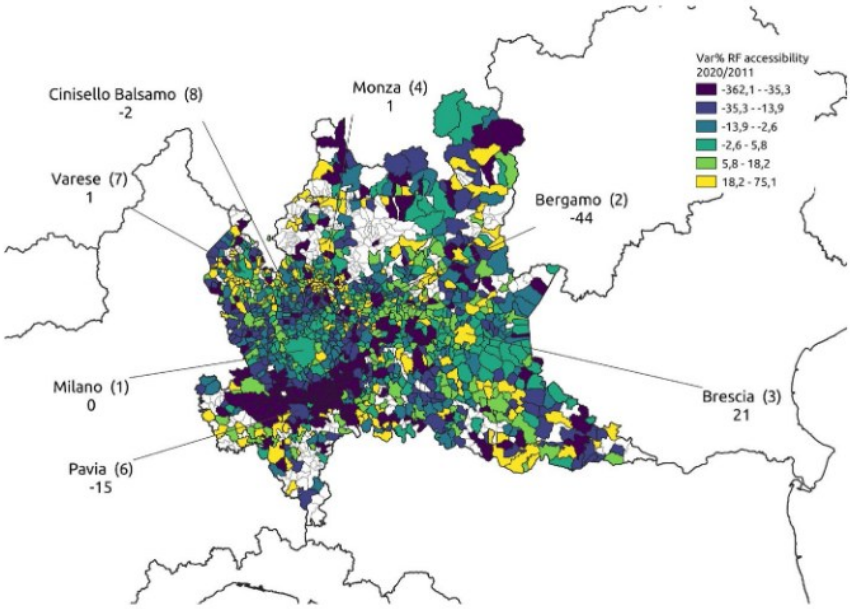
$$Acc_{2020} = ML(E_{2020}, Att_{2020}) \tag{10}$$

Mallien kykyä matkia HAM:ia arvioitiin Spearman korrelaatiolla, Kendallilla ja NDCG:llä (Normalized Discounted Cumulative Gain). Tulokset osoittivat, että RF oppii käytännössä täydellisesti HAM:in. Myös NN yltää korkeisiin arvoihin, mutta ei opi yhtä tarkasti kuin RF. Tämä tarkoittaa, että RF oppii siis paremmin HAM-hierarkian.

Taulukko 2:

| **Alue** | **ML** | **Spearman** | **Kendal** | **NDCG** |
|---|---|---|---|---|
| Lombardia | NN | 0,87 | 0,69 | 0,96 |
| Lombardia | RF | 1,00 | 1,00 | 1,00 |
| Emilia R. | NN | 0,96 | 0.83 | 0,99 |
| Emilia R | RF | 0,99 | 0,94 | 1,00 |

Tutkimus osoitti, että koneoppimista, etenkin kokonaisuusmetodi RF:ää, voidaan käyttää replikoimaan Hansenin saavutettavuusmallin käyttäytymistä ja arvioimaan saavutettavuutta vuosina, jolloin virtausdatat puuttuvat.




Kuva 6: Lombardian 2011 ja 2022 ennustettu variaatio saavutettavuudessa vuosien 2011 ja 2020 välillä (Hadjidimitriou, 2025).

# 5 Johtopäätökset

Kandidaatin työssä oli kolme tutkimuskysymystä: i) Millaisia koneoppimismenetelmiä on sovellettu saavutettavuustutkimukseen, ii) millaisia mahdollisuuksia ja haasteita niiden käyttöön liittyy ja iii) millaisia ongelmia niillä on mahdollista ratkaista tai yritetty ratkaista. Luvussa neljä käsiteltyjen tutkimusten perusteella koneoppiminen on osoittautunut tehokkaaksi tavaksi mallintaa ja ennustaa saavutettavuutta. Ohjattu oppiminen ja ohjaamaton oppiminen toimivat molemmat saavutettavuuden mallintamisessa.

## 5.1 Tulokset ja havaintoja kirjallisuudesta

Vastauksena tutkimuskysymykseen i) ja iii) tutkimusten perusteella sekä ohjaamattoman että ohjatun oppimisen algoritmit toimivat saavutettavuuden mallintamisessa. Ohjaamattoman ja ohjatun oppimisen yhdistämistä – niin sanottua hybridimallia voidaan käyttää saavutettavuusluokkien tutkimiseen ja ennustamiseen. Ensin jaotellaan alueet eri saavutettavuustasoihin jollakin ohjaamattoman oppimisen menetelmällä, esimerkiksi klusteroimalla. Sen jälkeen käytetään näitä klustereita datana, jolla koulutetaan jokin ohjatun oppimisen malli, esimerkiksi KNN. Näin voidaan automaattisesti luokitella hyvän ja huonon saavutettavuuden osia kaupungista.

Koneoppimista voidaan myös käyttää tutkimaan ja tunnistamaan etnisiä, sosiodemografisia ja taloudellisia tekijöitä, jotka vaikuttavat saavutettavuuteen. Tämä voidaan tehdä laskemalla saavutettavuus gravitaatiopotentiaalimallilla ja sen jälkeen käyttämällä ohjatun oppimisen koneoppimismenetelmää, tarkemmin niin sanottua imputoitua lassoa. Imputoidussa lassossa RF-algoritmi imputoi dataa, joka puuttuu, muuttujiin ja tämän jälkeen lassoa käytetään valitsemaan relevantit muuttujat.

ML-metodeilla voidaan toistaa HAM:in käyttäytymistä ja ennustaa saavutettavuutta, kun virtausdatat puuttuvat. Toisin sanoen ML voi extrapoloida ja toistaa HAM:in ja täten löytää datasta piilevän teorian, joka linkittyy SIM:iin. HAM-ML toimii siten, että luodaan ensin HAM, joka perustuu doubly-constrained SIM-kalibrointiin. Tämä Kalibroitu HAM toimii sitten ML-mallin kohdemallina. Tämän jälkeen koulutetaan ohjatun oppimisen algoritmi, esimerkiksi RF tai NN, jonka tehtävänä on tulevaisuudessa ennustaa saavutettavuutta, kun esimerkiksi virtausdata on puutteellista. Tulokset osoittivat, että RF toimii parhaiten tai ainakin paremmin, kun NN tässä oppimisskenaariossa.

Mielenkiintoinen – mutta ei kovin yllättävä – havainto tutkimuksista on, että kokonaisuusmetodit olivat hyvin toimivia koneoppimismenetelmiä

saavutettavuuden mallintamisessa. Kokonaisuusmetodit perustuvat siihen, että yhdistetään monta yksinkertaista heikkoa oppijaa (engl. weak learner) tuottamaan vahva oppija (engl. strong learner). Bootstrap-aggregation (Bagging) tekniikat, erityisesti satunnaismetsä, tuottivat hyviä tuloksia tutkimuksissa. Niiden toimivuutta muissakin saavutettavuuteen liittyvissä tutkimuksissa voisi tutkia.

Kokonaisuusmetodien toimivuus saavutettavuuden mallintamisesssa tukee myös kokonaisuusoppimisen teoriaa, jonka mukaan useiden heikkojen oppijoiden yhdistelmä voi muodostaa vahvasti yleistävän mallin. Myös muiden kokonaisuusmetodien kuten boostingin toimivuutta voitaisiin tutkia tulevaisuuden saavutettavuustutkimuksissa.

## 5.2 Haasteet ja kehityssuunnat

Tutkimuskysymykseen ii) liittyen koneoppimisen käyttö ei ole kuitenkaan haasteetonta. Suurimpia haasteita ovat datan laatu ja saatavuus. Vaikka algoritmi oppisi datan erittäin hyvin, jos data on virheellistä tai puutteellista tai muuten heikkolaatuista, algoritmi tuottaa hyödyttömiä tuloksia. Toinen haaste on mallien tulkittavuus ja niiden hyväksyttävyys päätöksenteossa. Kunnat ja hyvinvointialueet tekevät päätöksiä, joiden tulee olla perusteltavissa ja läpinäkyviä kansalaisille. Vaikka monimutkaiset mallit, kuten neuroverkot ja satunnaismetsät, voivat tarjota hyvää ennustetarkkuutta, ne näyttäytyvät helposti “mustina laatikoina”, joiden tuottamia tuloksia on vaikea selittää. Tämä voi rajoittaa ML-mallien hyödyntämistä esimerkiksi koulu- tai terveyskeskusverkkoa koskevissa poliittisissa päätöksissä. Siksi jatkokehityksessä olisi tärkeää yhdistää saavutettavuusmallinnukseen selitettävyysmenetelmiä tai suosia tarkoituksella yksinkertaisempia malleja, jotka ovat helpommin kommunikoitavissa.

## 5.3 Potentiaali käytännön sovelluksilla

Suomessa koneoppimismenetelmillä on huomattava potentiaali erityisesti sosiaali- ja terveyspalveluiden, joukkoliikenteen ja koulutusverkkojen saavutettavuuden arvioinnissa. Ohjaamattomiin menetelmiin perustuva klusterointi yhdistettynä ohjattuun oppimiseen – kuten Khosravi Khosravi ym. (2022) tutkimuksessa – voisi auttaa luokittelemaan Suomen kaupunki- ja maaseutualueet saavutettavuusluokkiin (”erittäin hyvä”–”erittäin heikko”). Näitä luokkia voitaisiin hyödyntää skenaarioanalyyseissa, joissa arvioidaan, miten yksittäisen terveyskeskuksen sulkeminen tai uuden palvelupisteen avaaminen muuttaa eri väestöryhmien saavutettavuutta.

# Lähteet


Alpaydin, E. (2020) *Introduction to Machine Learning*. Cambridge, MA: The MIT Press. ISBN 978-0262043793.

Burns, L.D. (1979) *Transportation, Temporal and Spatial Components of Accessibility*. Lexington/Toronto: Lexington Books.

Casali, Y., Nazli, Y.A. & Comes, T. (2022) 'Machine Learning for Spatial Anal-yses in Urban Areas: a Scopinf Review', *Sustainable Cities and Society*, 85, 104050. Saatavissa: https://doi-org.libproxy.aalto.fi/10.1016/j.scs.2022.104050

Dalvi, M.Q. & Martin, K.M. (1976) 'The measurement of accessibility: some premilinary results', *Transportation*, 5, pp. 17–42.

Fotheringham, A.S. & O'Kelly, M.E. (1989) *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer Academic Publishers. ISBN 0-7923-0021-1.

Geurs, K.T. & van Wee, B. (2004) 'Accessibility evaluation of land-use and transport strategies: review and research directions', *Journal of Transport Geography*, 12, pp. 127–140. Saatavissa: https://doi.org/10.1016/j.jtrangeo.2003.10.005

Hadjidimitriou, N.S., Reggiani, A., Östh, J. & Mamei, M. (2025) 'Hansen's Accessibility Theory and Machine Learning: a Potential Merger', *Networks and Spatial Economics*. Saatavissa: https://doi.org/10.1007/s11067-025-09674-2

Handy, S.L. & Niemeier, D.A. (1997) 'Measuring accessibility: an exploration of issues and alternatives', *Environment and Planning A*, 29, pp. 1175–1194.

Hansen, S. (1972) 'Utility, accessibility and entropy in spatial modelling', *The Swedish Journal of Economics*, pp. 35–44.

Hansen, W.G. (1959) 'How Accessibility Shapes Land Use', *Journal of the American Planning Association*, 25, pp. 73–76. Saatavissa: https://doi.org/10.1080/01944365908978307

Hägerstrand, T. (1970) 'What about people in regional science?', *Papers of the Regional Science Association*, 24, pp. 7–21.

Khosravi Kazazi, A., Amiri, F., Rahmani, Y., Samouei, R. & Rabiei-Dastjerdi, H. (2022) 'A New Hybrid Model for Mapping Spatial Accessibility to Healthcare Services Using Machine Learning Methods', *Sustainability*, 14, 14106. Saatavissa: https://doi.org/10.3390/su142114106

Kumar, S. & Bhatnagar, V. (2022) 'A Review of Regression Models in Machine Learning', *Journal of Intelligent Systems and Computing*.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018) *Foundations of Machine Learning*. 2nd edn. Cambridge, MA: The MIT Press. ISBN 978-0262039406.

Reddy, B.P., O'Neill, S. & O'Neill, C. (2022) 'Explaining spatial accessibility to high-quality nursing home care in the US using machine learning', *Spatial and Spatio-Temporal Epidemiology*, 41, 100503. Saatavissa: https://doi.org/10.1016/j.sste.2022.100503

Salze, P., Banos, A., Oppert, J., Charreire, H. & Casey, R. (2011) 'Estimating spatial accessibility to facilities on the regional scale: an extended commuting-based interaction potential model', *International Journal of Health Geographics*. Saatavissa: https://doi.org/10.1186/1476-072X-10-2

Wilson, A. (1967) 'A statistical theory of spatial distribution models', *Transportation Research*, 1(3), pp. 253–269.

Wilson, A. (1970) *Entropy in urban and regional modelling*. London: Pion. (Re-published by Routledge, Abingdon.)

Wilson, A. (2021) 'Entropy in urban and regional modelling', in *Handbook on Entropy, Complexity and Spatial Dynamics*. Cheltenham: Edward Elgar Publishing, pp. 20–31.