TURUN YLIOPISTO

Tekoälyn algoritmit MAA11 kurssille

Ongelmalähtöinen tehtäväpaketti lukio-opetukseen.

Matematiikan didaktinen
pro gradu -tutkielma

Laatija:
Kristian Juselius

23.5.2025
Turku

Pro gradu -tutkielma

Tutkinto-ohjelma, oppiaine: Matematiikan opettaja, matematiikka
Tekijä: Kristian Juselius
Otsikko: Tekoälyn algoritmit MAA11 kurssille
Ohjaajat: prof. Ion Petre, prof. Vesa Halava
Sivumäärä: 47 sivua
Päivämäärä: 23.5.2025

Tässä pro gradu -tutkielmassa esitellään tehtäväpaketti Lukion opetussuunnitelman 2019 mukaisen pitkän matematiikan 11 kurssille. Tehtäväpaketin keskeiset pedagogiset tavoitteet ovat esitellä tekoälyn luonne matemaattisena algoritmina sekä koneoppimisen merkitys yhteiskunnassa. Tehtävien aiheet ovat järjestyksessä suurdata, lineaarinen regressio, logistinen regressio, klusterointi sekä suuret kielimallit, ja jokaisesta tehtävästä esitellään kaksi lähestymistapaa sen mukaan, paljonko tähän halutaan käyttää aikaa. Tehtäväpaketti perustuu Euroopan unionin ja Yhdistyneiden kansakuntien kasvatus-, tiede ja kulttuurijärjestön (UNESCO) suosituksiin tekoälylukutaidon kehittämisestä.

Avainsanat: tekoälyopetus, matematiikka, lukio, tehtäväpaketti, ongelmalähtöinen opetus.

Sisällysluettelo

1 Johdanto

Tekoälyn käyttö on kasvanut räjähdysmäisesti 2010-luvulla niin arjessa kuin työelämässä. Käyttäjälle ei enää välttämättä ole selvää, miten hänen elämäänsä ohjailevat mallit toimivat, mistä ne saavat tietonsa eikä edes missä niitä käytetään (Kolari & Kallio, 2023). Tutkimukset ovatkin osoittaneet, että sekä median että keskivertoihmisen kuva tekoälystä on yhä vinoutunut kaupalliseen näkökulmaan, ja että tekoäly nähdään yhä useasti tieteisfantasiassa kuvattuna uhkana ihmiskunnalle (Slotte Dufva & Mertala, 2021). Vaikka tutkijat ovatkin toiseksi yleisin asiantuntijaryhmä tekoälyaiheisissa artikkeleissa 27 % osuudellaan ja vaikka 48 % ihmisistä uskoo tekoälyn helpottavan elämäänsä, edelleen yli kolmannes asiantuntijoista oli toimitusjohtajia, tai vastaavia kaupallisia toimijoita, ja 30 % ihmisistä uskoi kuvan tietoisen tekoälyn valloittavan ihmiskunnan (Slotte Dufva & Mertala, 2021).

Tekoälyaiheinen koulutus on selvästi tarpeen, mutta lainsäädäntö ja koulujärjestelmä ovat kuitenkin jäljessä tämän teknologian kanssa. Euroopan Unioni on julkaissut tekoälysäädöksen, joka velvoittaa avoimuuteen koneoppimisen ja sen kouluttamiseen käytetyn datan kanssa (European Union, 2024). Säädös myös kieltää suoraan tekoälyn soveltamisen ihmisoikeuksia uhkaavasti, esimerkiksi sosiaalisen pisteyttämisen (European Union, 2024). Tämän säädöksen ensimmäiset osat kielloista ja tekoälylukutaidoista tulivat voimaan helmikuussa 2025, ja vaikka ne on osoitettu yrityspuolelle, Helsingin kaupunki siteerasi kyseistä säädöstä tekoälysovellukset kieltävässä viestissään. Viesti oli osoitettu kaikille kaupungissa työskenteleville opettajille, ja se kielsi kaikkien tekoälysovellusten käytön koulutuksessa. Kaupungin edustajan mukaan viestin sisältö ei kuitenkaan vastannut tarkoitusta, vaan tarkoituksena oli kieltää Microsoft Copilot (Kymäläinen, 2025).

Euroopan unionin säädöksessä mainittu tekoälylukutaito ja sen kehitys osana kasvatusta on vasta aluillaan, mutta ensiaskeleet tekoälystä opettamiseen on jo otettu. Yhdistyneiden kansakuntien kasvatus-, tiede- ja kulttuuriorganisaatio UNESCO on julkaissut lausunnon tekoälykoulutuksessa, ja osana lausuntoa ehdotetaan opettajien tekoälyosaamisen kehittämistä, tekoälyn hyödyntämistä opetuksessa sekä kokeiluja uusiin opetusmenetelmiin aiheesta. Tekoälyä kehotetaan käyttämään työkaluna sekä opetuksessa, että oppimisessa. Esiin nostetaan kuitenkin opetuksessa käytettävien ohjelmien riskit – opetuksessa tulee huomioida niin tietoturva kuin etiikka ja ympäristövaikutukset.

Vaikka koneoppiminen on ollut tämän vuosituhannen suurimpia teknologisia läpimurtoja, matematiikka, johon koneoppiminen rakentuu, juontaa juurensa 1900-luvun puoleen väliin ja todelllinen läpimurto alalla ovatkin olleet suurdatan saatavuus internetin ansiosta ja tietokoneiden laskentatehon kasvaminen. Tekoäly saattaa muodostaa ihmisen ymmärryksen ylittävän kokonaisuuden, mutta sen perustoiminta on perimmiltään numeerista analyysiä, lineaarialgebraa ja funktion sovittamista datajoukkoon. Tämän pro gradu -tutkielman tarkoitus on luoda pohjaa tekoälyopetukselle tämän suhteellisen yksinkertaisen matemaattisen taustarakenteen avulla unohtamatta kuitenkaan tämän teknologian merkittäviä yhteiskunnallisia vaikutuksia.

1.1 Tekoäly kasvatuksessa (UNESCO)

Konkreettista tutkimusta tekoälystä kansallisen koulutuksen osana on vielä toistaiseksi hyvin vähän. UNESCO on kuitenkin julkaissut vuosina 2022–2024 useita tekoälykoulutuksen selvityksiä ja suuntaa antavia ohjeita. Näistä tärkeimpiä ovat 2022 julkaistu kartoitus tekoälylukujärjestyksistä jäsenmaissa, sekä ohjeet tekoälytietotaidoista opettajalle ja oppilaille vuoden 2024 Pekingin suurkokouksen jälkeen. Dokumenttien tarkoitus on rohkaista jatkotutkimukseen ja innovaatioihin, eivätkä ne anna täsmällisiä ohjeita tai rajoja tekoälykoulutukseen. Tämä johtuu siitä, miten uusi tekoälykoulutuksen osa-alue on. Edellä mainitut kolme dokumenttia ovat kuitenkin olennaisia tehtäväpaketin tavoitteita ja opetusmenetelmiä ajatellen.

UNESCOn 2022 julkaisemasta kartoituksesta selviää, että valtiotasolla käytettäviä tekoälykoulutuksen opetussuunnitelmia on ollut vuonna 2021 vain 11 valtiossa, eikä niidenkään tehokkuudesta ole vielä merkittävästi dataa. Suurin osa tekoälyopetussuunnitelmista on suunnattu yläaste- ja lukioikäisille, ja ne yleensä kuuluvat osaksi laajempaa tietoteknistä opetussuunnitelmaa. Serbia on ollut johtava kehittäjä viidellä opetussuunnitelmalla, joista kolme on ollut dokumentin julkaisuhetkellä kehityksessä. Matematiikka ei ole yleisesti lähtökohtana opetuksessa, vaan suurin osa opetussuunnitelmista kuuluu tietotekniikkamoduuleihin. Kartoitus on suunnattu enemmän koulutuksen organisoijille, kuin opetussuunnitelman kehittämisen ohjeeksi, mutta tämän tutkielman kannalta hyödyllisiä tuloksia ovat tekoälykoulutuksen suositellut tavoitteet ja menetelmät.

> *Recommendation 7.1: In low-resource contexts, curricula can focus on areas such as understanding AI, recognizing AI applications in everyday life, reflecting on social impacts, and engaging design thinking through paper prototypes or product redesign exercises.*
>
> *Recommendation 8.1: AI curriculum developers should consider leveraging innovative pedagogies to create interdisciplinary opportunities to solve real-life challenges faced by students and their communities, as a way to build skills in critical thinking, entrepreneurship, communication and teamwork.*
>
> *Recommendation 9.1: Curriculum development should focus on learning outcomes and the application of AI principles and processes rather than the ability to use specific platforms, devices or products. Where possible, curricula should engage a wide range of different technologies.*

Tässä tutkielmassa laadittu tehtäväpaketti on rajoitettu näiden suositusten mukaan käsittelemään tekoälyn matemaattista pohjaa sekä yhteiskunnallisia vaikutuksia, tehtävät ovat ongelmalähtöisiä ja tekoälyn leimaamista tiettyihin brändeihin vältetään aktiivisesti. Tällä leimaamisella ei kuitenkaan tarkoiteta tiettyjen brändien tai teknologioiden ottamista esimerkiksi tai työkaluksi osana tehtävää (UNESCO, 2022).

Vuoden 2022 kartoituksen lisäksi UNESCO on julkaissut vuonna 2024 tekoälytaidon puitteet (*AI Competency Framework)* sekä opettajalle, että oppilaalle. Toisin kuin 2022 kartoitus, nämä puitteet antavat suoraan ohjeita ja tavoitteita, joita tekoälykoulutuksessa tulisi ottaa huomioon. Opettajille ja oppilaille suunnattujen puitteiden yhteisiä teemoja ovat ihmiskeskeisyys, etiikka ja elämänmittaisen oppimisen tukeminen, vaikka jälkimmäiset ohjeistavat enemmän opettajien koulutukseen ja ammattitaidon kehittämiseen. Vaikka oppilaiden tekoälytaitojen puitteet ovat tämän tutkielman kannalta olennaisemmat, opettajien puitteissa ohjeistetaan tekoälyä ympäröivän kohun vaimentamisesta (*debunking the hype surrounding AI*) (Miao & Cukurova, 2024). Tämä suositus korostuu laaditussa tehtäväpaketissa, sillä yksi keskeisimmistä tavoitteista on tekoälyn ja koneoppimisen liittäminen algoritmeihin ja lineaarialgebraan, sekä algoritmien ’’yksinkertaisuuden’’ paljastaminen.

Tekoälypuitteet oppilaille sisältää enemmän ohjeita kurssien ja tehtävien suunnitteluun. Ne taas kehottavat myös tekoälyn sosiaalisista vaikutuksista ja etiikasta puhumiseen. Valistaminen datan alkuperästä nostetaan myös esille useita kertoja. Ohjeissa on myös taulukoituna tekoälyn opetuksessa huomioitavia keskeisiä tavoitteita, sekä se miten näitä voidaan toteuttaa. Nämä ovat taulukoitu ’’ymmärtää, soveltaa ja osaa luoda’’ asteikolla ikään kuin supistetun Bloomin taksonomian mukaan sekä seuraavien aihealueiden mukaisesti: ihmiskeskeinen lähestyminen, tekoälyn etiikka, tekoälyn tekniikat ja sovellukset sekä tekoälyjärjestelmien suunnittelu.

1.2 Sanasto

Koska tekoälyn ja koneoppimisen sanasto ja terminologia ei ole vielä vakiintunutta, tässä kappaleessa esitellään tutkielmassa käytetyt termit, niiden alkuperä ja tarkoitus. Pääasiallisena lähteenä on käytetty kirjaa Tekoäly 123 sekä Suomen Tilastoseuran sanakirjaa (Kolari & Kallio, 2023) (Alho, Arjas, Läärä, & Pere, 2023). Poikkeuksena on tutkielmassa käytetty käännös ''leima'' sanasta ''*label*'', jolle ei löytynyt tutkielman kirjoitushetkellä vakiintunutta ja pedagogisesti kuvaavaa käännöstä. Vaihtoehtoisia käännöksiä ovat ''nimiöity data'' (*labeled data*) tai ''merkittävä/merkitty piirre.'' Alla listaus tutkielmassa käytetystä terminologiasta

Harjoitusdata (Training data)	Mallin kouluttamiseen käytettävä datajoukko
Koneoppiminen (Machine learning)	Tekoäly, joka kykenee itsenäiseen oppimiseen.
Leima (Label)	Ohjatussa koneoppimisessa käytetty merkitty piirre, joka kertoo mikä arvo harjoitusdatasta haluttaisiin saada.
Lineaarinen regressio (Linear regression)	Yksinkertainen ohjattu koneoppimismenetelmä, jossa datasta pyritään ennustamaan haluttu ominaisuus muiden piirteiden lineaarikuvauksena.
Logistinen regressio (Logistic regression)	Yksinkertainen ohjattu koneoppimismenetelmä, jossa data jaetaan kahteen luokkaan logistisen funktion avulla.
Luokittelualgoritmi (Classification algorithm)	Algoritmi, joka jakaa sille syötetyn datan ennalta määrättyihin luokkiin.
Neuroverkko (neural network)	koneoppimismalli, joka yhdistää lineaarisilla kuvauksilla neuroneita ja näiden epälineaarisia aktivointifunktioita useassa kerroksissa.
Ohjaamaton koneoppiminen (Unsupervised machine learning)	Koneoppimisen muoto, jossa datalla ei ole leimoja, eikä sitä voida ohjata oikeaan suuntaan virhe- tai tappiofunktion avulla.
Ohjattu koneoppiminen (Supervised machine learning)	Koneoppimisen muoto, jossa algoritmia ohjataan oikeisiin vastauksiin rankaisemalla sitä vääristä tai epätarkoista vastauksista.
Piirre (Feature)	Datavektorin komponentti, joka kertoo nimetyn ominaisuuden, esimerkiksi henkilön pituus, paino jne.

Ryvästys (Clustering)	Ohjaamaton koneoppimismenetelmä, jossa datasta yritetään löytää säännönmukaisuuksia, eli ryppäitä. Vaihtoehtoinen suomennos: klusterointi.
Sanavektori (Word vector**)**	Ohjaamattomalla koneoppimisella luotu vektorimuotoinen muoto luonnollisen kielen sanalle, jossa samassa kontekstissa esiintyvät sanat sijaitsevat lähellä toisiaan.
Suurdata (Big data)	Nimensä mukaan suuri joukko dataa, jossa usein kymmeniä tuhansia pisteitä ja kymmeniä tai satoja piirteitä. Sisältää yleensä sekalaista ja epäjohdonmukaista tietoa ja ei yleensä pystytä käsittelemään perinteisin tietojenkäsittelymenetelmin.
Suuri kielimalli (Large Language Model/LLM)	Moniosainen koneoppimismalli, joka kykenee analysoimaan ja tuottamaan luonnollista kieltä.
Tekoäly (Artificial intelligence):	Tietokoneohjelma, joka kykenee imitoimaan älyllistä tai ihmistä muistuttavaa toimintaa.
Testidata (Test data)	Datajoukko, jolla lopullisen koneoppimismallin toiminta testataan. Tärkeää erottaa muusta prosessista, jottei mallia rakenneta vahingossa testidataa varten.
Validointidata (Validation data)	Datajoukko, jolla mallin toiminta testataan koulutuksen jälkeen. Antaa mahdollisuuden säätää mallin ominaisuuksia ja testata useita malleja ennen testidatan käyttöä.

2 Tavoitteet ja rakenne

Tehtäväpaketin tavoitteet ja rakenteet perustuvat edellä mainittuihin UNESCOn ja Euroopan unionin suosituksiin. Keskeisiä teemoja tehtäväpaketissa ovat tekoälyn selittäminen matematiikan kautta, eettisten kysymysten huomioiminen sekä ongelmalähtöinen lähestymistapa. Kurssin MAA11 ja lukion opetussuunnitelma tavoitteet on myös pyritty huomioimaan. Kurssin laajuuden takia tehtäväpaketti ei saa olla liian työläs ja sen tulisi olla luonteva lisä kurssiin, eikä korvata olemassa olevia materiaaleja. Ihanteellisesti tehtäväpaketin sisältöä käytetään soveltavina tehtävinä kurssin loppuosassa. Tässä tutkielmassa ei esitetä ohjelmointia vaativia tehtäviä, mutta ne tukisivatt sekä tehtäväpaketin että kurssin tavoitteita, joten niiden käyttö ja kehittäminen on suositeltavaa (Opetushallitus, 2019). Tehtäväpaketti ei perustu ns. ohjelmointitehtäviin, jotta sen käyttöönotto olisi helppoa opettajan taitotasosta huolimatta ja jotta tehtäväpaketin ottaminen mukaan opetukseen vaatisi koululta mahdollisimman vähän resursseja. Tämä on myös linjassa UNESCOn suositusten kanssa tilanteissa, joissa tekoälylukutaitokoulutukseen ei ole paljoa resursseja (UNESCO, 2022).

2.1 Tavoitteet

Tehtäväpaketin keskeisimmät tavoitteet ovat tekoälyn matemaattiset perusteet, suurten kielimallien (*large language models*) käyttö sekä tekoälyn etiikka ja sosiaaliset vaikutukset. Matemaattiset perusteet ja suurten kielimallien käyttö on jaoteltu aihealueittain omiin tehtäviin, kun taas etiikka ja sosiaaliset vaikutukset on sisällytetty näihin tehtäviin ns. ”upotetun etiikan” periaatteiden mukaan (*embedded ethics).* Tässä lähestymistavassa eettiset kysymykset pidetään olennaisena osana koko prosessia tekoälyn käytössä ja kehittämisessä sen sijaan, että niitä ajateltaisiin koneoppimismallien kehittämisen erillisenä osana jälkikäteen (Grosz, et al., 2018). Tämä lähestymistapa sopii laadittuun tehtäväpakettiin erityisen hyvin sillä se pitää pääpainon matematiikassa, mutta antaa mahdollisuuden pohtia pinnalla olevaa keskustelua koneoppimisen käytöstä. Eettisten kysymysten käsitteleminen tuo tehtäväpakettiin myös lukion opetussuunnitelman kannustamaa monialaista osaamista (Opetushallitus, 2019).

Tehtäväpaketin tavoitteet voidaan jakaa aiheittain matemaattisiin tavoitteisiin sekä eettisiin tavoitteisiin ja oppimisen syvyys uudistetun Bloomin taksonomian mukaan asteikolla muistaminen, ymmärtäminen, soveltaminen, analysoiminen ja arvioiminen (Anderson, et al., 2014). Syvintä Bloomin taksonomian tasoa ’’luominen’’ ei käytetä, sillä se ei olisi realistista tehtäväpaketin laajuuden ja opetuksen resurssien puitteissa, vaikka ohjelmoinnissa harrastuneelle oppilaalle ei välttämättä olisi mahdotonta päästä tälle tasolle, esimerkiksi luomalla oma yksinkertainen koneoppimismalli esimerkiksi python-ohjelmointikielellä. Muistamisella tarkoitetaan edellä mainitulla em. asteikossa pinnallista tietoa, jossa oppilas osaa palauttaa opittuja faktoja. Ymmärtämisen tasolla näitä muistettuja tietoja osataan vertailla, järjestellä ja kuvailla (Anderson, et al., 2014). Soveltamisella tarkoitetaan, että oppilas pystyy käyttämään aiempia tietoja uusissa tilanteissa uusilla tavoilla (Anderson, et al., 2014).

Analysoinnilla tarkoitetaan kykyä päätellä annetusta tiedosta motiiveja ja syitä sekä perustella näitä päätelmiä faktoilla. Arviointi-tason ero analysointiin on se, että arvioinnissa päätelmien pohjalta muodostetaan perusteltuja mielipiteitä (Anderson, et al., 2014).

Vektorilaskennan perusteet ovat olennainen pohjatieto tehtäväpaketille, ja tätä osaamista syvennetään tehtäväpaketissa yli kolmeulotteisiin vektoreihin. Oppilaan tulisi muistaa datan olevan moniulotteisessa vektorimuodossa, jossa yksi datapiste muodostaa n-ulotteisen vektorin. Oppilaan tulisi myös ymmärtää datan käsittelyn merkitys koneoppimisessa, sekä pystyä analysoimaan datan selkeitä ongelmakohtia ja arvioimaan päällisin puolin datan sopivuutta malliin. Oppilaan ei tarvitse osata varsinaisia tilastollisia keinoja datan arviointiin, mutta hänellä tulisi olla tehtäväpaketin jälkeen intuitiivinen käsitys, miten data voi vinouttaa mallia esimerkiksi liian pienen otoksen tai huonosti valikoitujen ominaisuuksien kautta. Tämän tavoitteen on tarkoitus kehittää oppilaan kriittistä ajattelukykyä tekoälyn toimintatavasta ja sen käytöstä päätösten teossa: tekoäly on koulutettu aina jollakin datalla, johon se perustaa päätelmänsä, eikä se pyri tietoisesti korjaamaan datasta aiheutuvia ongelmia, vaan tämä työ jää dataa käsitteleville ihmisille. Datan merkityksen tavoitteet ovat jatkuvana osana kaikkia tehtäväpaketin tehtäviä.

Ensimmäinen koneoppimiseen erikoistunut aihe on lineaarinen regressio. Vaikka aihe yksinkertaistetaan kaksiulotteisen suoran sovittamiseen, oppilaan tulisi muistaa, että todellisissa tilanteissa, kyseessä olisi n-ulotteinen suora. Virhefunktio ja gradienttimenetelmä ovat myös käsitteet, joita oppilaan ei tarvitse ymmärtää muistamista korkeammilla osaamisen tasoilla, sillä ne niiden täsmällinen matematiikka vaatii lineaarialgebraa ja analyysiä, joita lukion opetussuunnitelmaan ei kuulu. Ne ovat kuitenkin aiheita, jotka ovat suhteellisen helppo ymmärtää kaksi- tai kolmeulotteisissa tilanteissa. Lineaarikuvaus, virhefunktion muodostaminen, sekä parametrien päivitys gradienttimenetelmällä muodostavat koneoppimisen perusalgoritmin, mikä on syytä ymmärtää perusosana laajempia koneoppimismalleja. Tämä algoritmi myös liittää koneoppimisen MAA11 kurssiin. Oppilaan tulisi pystyä soveltamaan tietoja yksinkertaisten ohjelmointitehtävän yhteydessä, esimerkiksi tarkastamaan ohjelmakoodin muuttujia ja hänen tulee kyetä arvioimaan lineaarisen regression soveltuvuutta erilaisisissa tehtävissä, esimerkiksi kykeneekö se ottamaan huomioon kaikkia muuttujia, tai onko ongelma edes lineaarisesti kuvattavissa. Vaikka lineaarista regressiota käsitellään laaditussa tehtäväpaketissa vain yhden tehtävän verran, sen aiheet toistuvat muissakin tehtävissä.

Toinen laadituista tehtävistä käsittelee luokittelutehtävää. Oppilaalle esitellään muistettavalla tasolla logistinen funktio, sekä logaritmivirhefunktio. Näistä oppilaalle riittää muistaa niiden pääpiirteet ja syy niiden käyttöön: logistisen funktion muoto ja sijoittuminen välille (0,1) ja logaritmivirheen tarkoitus rankaista mahdollisimman paljon väärää vastausta ja palkita oikeaa. Opiskelijan tulisi ymmärtää luokittelumallin yhteys lineaariseen regressioon, sekä monen luokan yksinkertaistuminen moneen yksinkertaiseen luokitteluun. Soveltamistaso jää matemaattisen monimutkaisuuden vuoksi logistisen funktion sovittamiseen annettuihin datapisteisiin lukiosta tutuilla työkaluilla. Analysoinnin ja arvioinnin tasoilta tavoitteena on vain osata arvioida luokittelumallin sopivuutta ongelman ratkaisuun,

esimerkiksi onko ongelma edes ratkaistavissa luokittelumallilla. Tavoitteet lineaarisen regression ja luokittelun kanssa ovat hyvin samanlaiset, sillä ne eivät poikkea merkittävästi algoritmeiltaan, ja luokittelualgoritmin matemaattiset ominaisuudet menevät monilta osin yli lukion opetussuunnitelman.

Laaditun tehtäväpaketin neljäs tehtävä sekä viimeinen koneoppimismenetelmä on ryvästys, joka on kerännyt huomiota etenkin sosiaalisessa mediassa, sillä käyttäjät jättävät jälkeensä massiivisia määriä dataa. Palvelun tuottajat ovat olleet jo pitkään kiinnostuneita kohdentamaan tämän datan avulla sisältöä käyttäjäryhmille (Alsayat & El-Sayed, 2016). Tarkkaa tietoa ryvästysmenetelmien käytöstä kaupallisesti ei kuitenkaan ole, sillä algoritmeja ei kirjoitushetkellä ole julkistettu. Oppilaan tulisi muistaa pääpiirtein hierarkkisen-, keskiarvollisen- ja jakaumallisen ryvästyksen erot sekä ymmärtää ryvästys ohjaamattomana koneoppimismenetelmänä ja tämän ero ohjattuihin menetelmiin. Tavoitteena soveltamisen suhteen on keskiarvoalgoritmin simulointi kaksiulotteisessa tapauksessa ja oppilaan tulisi pystyä arvioimaan ryvästysmenetelmien soveltuvuutta ja ajankohtaisuutta tekoälykeskustelussa, esimerkiksi koskien niiden käyttöä sosiaalisen median tai mainonnan algoritmeissa.

Laaditun tehtäväpaketin viimeinen aihe käsittelee suuret kielimallit sekä neuroverkot. Tästä aiheesta oppilaiden ei tarvitse osata matemaattista mallinnusta – tarvittava lineaarialgebra vaatii jo korkeakoulutason matematiikkaa. Oppilaan tulee muistaa käsitteellisesti sanat semanttisina vektoreina, eli sanavektoreina, sekä ymmärtää neuroverkkojen muodostuminen lukuisista yksinkertaisemmista koneoppimismalleista, joiden tuloksia arvioidaan ja painotetaan useilla tasoilla. Ylempiä oppimisen tasoja tässä aiheessa arvioidaan suurten kielimallien käyttötaidoilla. Oppilaan tulee osata soveltaa suuria kielimalleja, kuten Chat-GPT tai Microsoft Copilot tekoälyjä projektissa. Kielimallien antamia vastauksia tulee kuitenkin pystyä analysoimaan: Oliko tekoälytyökalun vastaus oikea, onko mahdollisille virheille selkeitä syitä ja voisiko virheet välttää käyttämällä erilaista komentoa. Näiden taitojen tarkoituksena on rakentaa oppilaan kykyä arvioida kriittisesti tekoälytyökalun käytön hyötyjä ja haittoja jokapäiväisissä projekteissa.

Kuten edellä on mainittu, tekoälyn eettisiä kysymyksiä käsitellään pitkin tehtäväpakettia ja näiden eettisten teemojen tehtävä on valmistaa oppilaat käsittelemään tekoälyn käytön yhteiskunnallisia vaikutuksia. Muistamistason tavoitteena on, että oppilas kykenee antamaan yksittäisiä esimerkkejä tekoälymallien käytöstä sekä niiden herättämästä kritiikistä. Nämä esimerkit voivat liittyä esimerkiksi spekuloituihin tekoälyalgoritmeihin sosiaalisessa mediassa, itse ajaviin autoihin tai rekrytointiprosessien automatisaatioihin. Oppilaan pitäisi myös ymmärtää miten tekoälyn käyttämä data heijastaa ja uusintaa yhteiskunnallisia arvoja, esimerkiksi miten rekrytointiprosessi voi suosia tiettyä ihmisryhmää lähihistorian syrjinnän ja segregaation takia. Tekoälymallien tuottamia ongelmia ja niiden syitä pitäisi pystyä myös arvioimaan esimerkiksi koulutusdatan laajuuden tai laadun kautta. Viimeisenä tehtäväpaketin keskeisenä tavoitteena on kehittää oppilaan kykyä arvioida tekoäly- ja koneoppimisten soveltuvuutta yhteiskunnallisiin tehtäviin ja ihmisiä koskevaan päätöksentekoon.

Taulukko 1: Tehtäväpaketin tavoitteet aiheittain

	Muistaa	Ymmärtää	Soveltaa	Analysoi ja arvioi
Suurdata	Datan muoto vektorina.	Datan käsittelyn merkitys koneoppimismal-leissa	Osaa rakentaa annetusta datasta vektorin ja käyttää tätä yksinkertaisissa algoritmeissa.	Analysoi: Datan soveltuvuus ja selkeät ongelmakohdat.
Lineaarinen regressio	käsitteet, lineaarinen kuvaus, virhefunktio ja gradienttimenetel-mä, sekä niiden merkityksen osana algoritmia.	Lineaarisen reg-ression algoritmin merkitys koneop-pimisen peruso-sana.	Osaa muodostaa vuokaavion lineaarisesta regressiosta.	Arvioi: lineaarisen regression soveltuvuus annettuun tehtävään.
Luokittelu	Logistinen funktio ja logaritmivirhe sekä niiden merkitys ja syy.	Yhteys lineaariseen regressioon ja monen luokan yksinkertaistumi-nen moneen kahden luokan luokitteluun	Osaa sovittaa logistisen funktion luokiteltaviin datapisteisiin.	Arvioi: luokittelumallin soveltuvuus annettuun tehtävään.
Ryvästys	Pääpiirteiset erot hierarkkisen-, keskiskiarvollisen- ja jakaumallisen ryvästyksen välillä.	Ohjatun ja ohjaamattoman koneoppimisen ero.	K-keskiarvomene-telmän simulointi kaksiulotteisessa tapauksessa ja vuokaavion laatiminen.	Analysoi: Ryvästysmene-telmien soveltuvuus annettuun tehtävään. Arvioi: Ryvästyksen mer-kitys sosiaalisessa mediassa ja mainonnassa.
Suuret kielimallit ja neuroverkot	Sanavektorit, eli sanojen määrittäminen semanttisina vektoreina.	Neuroverkkojen ja suurten kielimallien rakenne pelkistet-tynä monitasoisiin koneoppimsmal-leihin	Suurten kielimallien käyttö työkaluna projekteissa.	Analysoi: Tekoälylle syötetty-jen komentojen ja sen antamien vastausten laadun kriittinen tarkastelu. Arvioi: Tekoälyn käytön hyödyt ja haitat arkisissa projekteissa.
Etiikka	Esimerkit tekoälyn käytöstä ja kritiikistä.	Miten tekoäly ja sen käyttämä data heijastaa ja uusintaa yhteiskunnallisia arvoja?	Tekoälytyökalujen järkevä ja perusteltu käyttö.	Analysoi: Tekoälykeskustelun kriittinen tarkastelu. Arvioi: Tekoälyn käyttö ja merkitys yhteiskunnassa.

2.2 Rakenne

Tehtäväpaketti koostuu viidestä tehtävästä, jotka kukin käsittelevät yhtä koneoppimismenetelmää pois lukien ensimmäinen tehtävä, joka pohjustaa vektorilaskentaa ja käsitteistöä. Tehtävien aiheet ovat läpikäyntijärjestyksessä suurdata ja koneoppiminen, lineaarinen regressio, luokittelumenetelmä, ryvästys ja suuret kielimallit. Projektimuotoiset tehtävät ovat UNESCOn selvityksen mukaan osoittautuneet tehokkaimmiksi tehtävätyypeiksi tekoälykoulutuksessa, joten tehtävien taustalla on myös ajatus tukea ongelmalähtöistä opetusta (UNESCO, 2022). Onglemalähtöinen opetus on kerännyt paljon kannatusta ja positiivisia tutkimustuloksia on raportoitu niin kansallisesti Suomessa kuin maailmallakin (Tan, 2009). Tehtäväpakettia soveltaessaan opettajan olisi syytä muistaa, että ongelmalähtöinen oppiminen ei ole ainoastaan tehtävänannosta kiinni ja, jotta opetusmenetelmä olisi potentiaalisimmillaan, tehtäväpakettia pitäisi opettaa ongelmalähtöisin opetuskeinoin tunneilla ja niiden ulkopuolella (Poikela, Vuoskoski, & Kärnä, 2009). Tehtävien rakenne on sellainen, että ne voidaan antaa ryhmätehtävinä oppilaille itseopittavaksi ja purettavaksi opettajan johdolla, mutta johtuen tehtäväpaketin luonteesta lisänä olemassa olevalle kurssille, tämä sitoutuminen tiettyyn opetustapaan ei ole realistinen. Tehtävänannot on tämän takia annettu sekä ongelmalähtöisen opetustavan mukaan, että yksityiskohtaisempina perinteisinä tehtävänantoina, jotta tehtävät olisivat mahdollisimman käyttökelpoisia kurssin ajankäytön puitteissa.

Koska tekoälyn eettiset kysymykset ovat keskeinen osa kansainvälistä huolta tekoälykoulutuksessa, usko laajaan tekoälyyn on edelleen yleistä ja koska kaupalliset toimijat edustavat yhä suurta osaa tekoälykeskustelusta, eettisiä kysymyksiä ei tulisi ohittaa (UNESCO, 2019) (Slotte Dufva & Mertala, 2021). Tämän takia eettiset kysymykset on sisällytetty tehtäviin Harvardin yliopiston upotetun etiikan (*embedded ethics*) mukaan. Syyt tähän toteutukseen ovat ajan kohdentaminen kurssin MAA11 tavoitteisiin ja se, että vaikka upotetun etiikan käyttö on ollut vasta koekäytössä, tulokset koekäytössä ovat olleet lupaavia. Käytännössä upotettu etiikka tarkoittaa, että tehtävät on rakennettu käsittelemään ongelmia, joihin tekoälyä saatetaan oikeasti käyttää ja käyttäen oikeaa dataa, mikäli se on mahdollista (Grosz, et al., 2018). Tehtävissä saattaa olla myös konkreettinen tehtävänanto eettisiin kysymyksiin liittyen, mutta erillistä tehtävää tekoälyn etiikasta ei ole. Opettajan ohjaama keskustelu aiheesta on resurssien puitteissa tietenkin suositeltavaa.

Tehtäväpaketin ensimmäinen tehtävä käsittelee dataa. Tehtävässä esitetään data moniulotteisena vektorina ja pohjustetaan käsitteitä koneoppimiseen ja algoritmiikkaan. Tehtäväpaketin tekohetkellä käytetään vuoden 2019 opetussuunnitelmaan, jossa vektorit käsitellään enemmän kaksi- tai kolmeulotteisina geometrisinä siirtyminä, kuin matemaattisina objekteina, eikä käsitettä laajenneta käsittämään yli kolmannen ulottuvuuden eikä matriiseista puhuta ollenkaan. Matriisin käsitettä ei tässä tehtäväpaketissa tarvita, mutta datan muoto moniulotteisena vektorina on kuitenkin olennainen käsite koneoppimisen ja suurdatan ymmärtämisen kannalta. Tämän lisäksi oppilaalle esitellään koneoppimisen käsitteistöä ja historiaa pääpiirteittäin, esimerkiksi milloin koneoppimista on alettu tutkimaan ja mistä

sen suosio 2010-luvulla johtuu. Tehtävä kulminoituu datan merkitykseen koneoppimisessa ja tekoälymalleissa sekä sen runsaaseen saatavuuteen internetin aikakaudella.

Toinen tehtävä käsittelee lineaarista regressiota, joka on kenties yksinkertaisin ohjattu koneoppimismalli. Tehtävässä oppilas pääsee tutustumaan, miten suurdatajoukosta voidaan ennustaa tuloksia, esimerkiksi ennustaa asunnon myyntihinta sen koon ja sijainnin perusteella. Oppilaalle esitellään, miten datan vektorista voidaan tehdä yksinkertainen lineaarinen kuvaus, jonka parametrit määritetään algoritmillisesti. Osa lineaarisen regression toiminnasta vaatii numeerista analyysiä ja lineaarialgebraa, joka ei kuulu lukion oppimäärään, mutta algoritmin toiminta on ymmärrettävissä, ja piirrettävissä vuokaavioon ilman täsmällistä matemaattista käsittelyä.

Kolmas tehtävä jatkaa lineaarisen regression teemasta, mutta laajentaa saman mallin luokittelutehtävään. Tehtävän premissinä on käyttää logistiseen funktioon upotettua lineaarista kuvausta jakamaan kaksi datapistejoukkoa, ja ennustaa kumpaan joukkoon uusi piste kuuluisi. Esimerkki voisi olla vaikka, onko kasvain hyvän- vai pahanlaatuinen sen koon ja muodon perusteella tai miten kuvasta voidaan tunnistaa mikä kirjain siihen on kirjoitettu. Lineaarinen kuvaus ja virhefunktio laajennetaan logistiseen funktioon ja logaritmivirheeseen, joista oppilaiden pitäisi hahmottaa muoto kaksiulotteisessa tapauksessa ja tarkoitus niiden käyttöön. Algoritmillisesti luokittelumalli ja lineaarinen regressio ovat hyvin samanlaisia, joten nämä kaksi tehtävää ovat myös rakenteeltaan samankaltaiset: oppilaat tutkivat kaksiulotteisissa tilanteissa kuvaajan sovittamista dataan ja pohtivat miten hyvin kyseinen malli ennustaa arvon testidatalle.

Neljännessä tehtävässä käsitellään ryvästys ja ohjaamaton koneoppiminen. Oppilaat tutkivat miten datasta, jossa ei ole valmiita leimoja voidaan etsiä samankaltaisuuksia ja keskittymiä, eli niin kutsuttuja ryppäitä. Oppilaille esitellään hierarkkinen ryvästys, jossa datapisteitä yhdistellään lähimpiin pisteisiin, kunnes datalle on saatu haluttu muoto, jakaumapohjainen klusterointi, jossa dataan sovitetaan todennäköisyyskuvaajat maksimoiden datan todennäköisyys kuulua kyseisille jakaumille, sekä k-keskiarvomenetelmä, jossa datalle määritetään haluttu määrä keskiarvoja ja etsitään näiden keskiarvojen sijainnit. Näistä ryvästysmalleista täsmällisempi matematiikka avataan k-keskiarvomenetelmälle ja sitä simuloidaan kaksiulotteisessa tapauksessa. Oppilaat myös pohtivat teknologian merkitystä kaupallisessa käytössä ja esimerkiksi viihdesovellusten suositusalgoritmeissa.

Viimeisessä tehtävässä oppilaat tutustutetaan suuriin kielimalleihin ja neuroverkkoihin. Neuroverkkojen matemaattinen mallintaminen vaatii matriisilaskentaa, joka ei kuulu enää lukion oppimäärään, joten matemaattinen osa neuroverkoista jää käymättä. Ne esitetään kuitenkin monikerroksisina koneoppimismalleina, joissa esimerkiksi lineaarisia regressioita tehdään useassa eri tasossa. Oppilaalle esitellään myös käsite sanavektoreista, eli ohjaamattomasta koneoppimismenetelmästä, jossa luonnollisen kielen sanoista muodostetaan semanttisia vektoreita niiden merkityksen ja käytön kontekstin mukaan. Tehtävässä oppilaat saavat valita projektiaiheen, jonka toteutuksessa he käyttävät suurta kielimallia apunaan, sekä dokumentoivat ja arvioivat kyseisen tekoälytyökalun komentoja ja vastauksia mahdollisimman tarkasti. Tehtävän tarkoituksena on lähestyä

moderneja tekoälytyökaluja, kuten Chat-GPT ja Microsoft Copilot, kriittisestä näkökulmasta ja pohtia niiden käyttötarkoitusta, tehokkuutta ja tarvetta.

3 Tehtävä 1: Data

Tässä tehtävässä oppilaat tutustuvat suurdatan käsitteeseen ja merkitykseen koneoppimisessa. Tehtävässä havainnollistetaan suurdatan runsautta internetin aikakaudella sekä miten ilmeisen satunnainen ja merkityksetön data voi pitää sisällään arvokasta tietoa.

3.1 Tehtäväkohtaiset tavoitteet

Tämä tehtävä pohjustaa koneoppimisen lineaarialgebraa ja käsitteistöä, koska lukion opetussuunnitelmassa ei opeteta vektoreita kolmea suuremmissa ulottuvuuksissa tai vektoriavaruuden alkioina. Tämä käsitteellinen ero kaksi- tai kolmiulotteisen ’’siirtymän’’, jolla on suunta ja suuruus, ja vektoriavaruuden alkion välillä rajoittaa vektorien käyttöä datan käsittelyssä ja algoritmeissa. Oppilaan ei kuitenkaan tarvitse ymmärtää vektorin määritelmää n-ulotteisen lineaariavaruuden alkiona, vaan hänelle riittää muistaa, että vektorit voidaan laajentaa yli kolmeen ulottuvuuteen ja kullekin ulottuvuudelle voidaan osoittaa yksi datan piirre. Datalla, sen keräämisellä ja laadukkaalla käsittelyllä on myös suuri merkitys tekoälyssä ja koneoppimisessa eikä se ilmesty tyhjästä. On myös olennaista huomata, että vaikka tekoälyä ja koneoppimista pidetään poliittisesti ja eettisesti puolueettomana, data, jolla mallit on koulutettu ja tarkistettu eivät sitä välttämättä ole. Näihin tarkoituksiin käytetty data ei myöskään ole välttämättä julkisesti näkyvillä tai tarkasteltavissa. Oppilaille havainnollistetaan mitä on suurdata, mistä sitä kerätään ja miten sitä voi hyödyntää.

3.2 Materiaalit ja eriyttäminen

Opettajan tulee valmistua selittämään opiskelijoille käsitteellisesti vektorit useassa ulottuvuudessa. Tätä varten on syytä ensin irtautua yksikkövektoriesityksestä ja näyttää, että esimerkiksi

$$\text{vektori } 3i + 5j - k \text{ voidaan esittää } (3 \quad 5 \quad -1).$$

Koska yksikkövektoreiden symbolit eivät enää rajoita esitystä, mikään ei estä lisäämästä sulkujen sisään lisää komponentteja. Oppilaille on myös syytä täsmentää, että kolmen ulottuvuuden jälkeen nuoliesitys ei enää ole järkevä, sillä piirrokset esimerkiksi neliulotteisesta avaruudesta ovat jo huomattavan sekavia. Oppilaille on syytä näyttää myös muutama esimerkki, miten yhteen- ja vähennyslaskut sekä skalaaritulo toimivat täsmälleen samalla tavalla kuin kolmiulotteisessakin tilanteessa esimerkiksi yhteenlasku

$$\begin{aligned}(3 \quad 5 \quad -1 \quad -2) + (7 \quad -11 \quad 7 \quad -6) &= (3+7 \quad 5+(-11) \quad -1+7 \quad -2+(-6)) \\ &= (10 \quad -6 \quad 6 \quad -8)\end{aligned}$$

tai skalaaritulo

$$-3 \cdot (3 \quad 5 \quad -1 \quad -2) = (-3 \cdot 3 \quad -3 \cdot 5 \quad -3 \cdot (-1) \quad -3 \cdot (-2))$$
$$= (-9 \quad -15 \quad 3 \quad 6).$$

Tästä on luonteva siirtyä listoihin ohjelmoinnissa ja siten dataan, jossa yhdellä pisteellä voi olla monta piirrettä. Tästä oppilaille voi näyttää kuvan 1 mukaisen esimerkin, miten heidän koulustaan voidaan löytää useita eri lukuarvoja tai kategorisia piirteitä.

	A	B	C
1	nimi	Kerttulin lukio	
2	Sijainti	60,449521° N, 22,283247° E	
3	oppilasmäärä	600	
4	Henkilökunnan määrä	80	
5	Koulutyyppi	Lukio	
6	sisäänpääsyraja	8,75	
7	Erikoislinja(t)	urheilu, ICT	
8	Perustamisvuosi	2012	
9	YO-kokeiden tulosKA	5,2	
10	rakennuksen rakennusvuosi	1912	
11			

Kuva 1: Esimerkki tehtävän taulukosta.

3.2.1 Eriyttäminen

Alaspäin eriytettäessä oppilaille voidaan antaa yksinkertaisesti vähemmän dataa etsittäväksi tai opettaja voi antaa valmiiksi tehdyn taulukon kuvan 1 mukaan. Datan vähentäminen säästää aikaa, eikä vaikuta merkittävästi tehtävän tavoitteiden toteutumiseen, mutta valmiin taulukon käyttäminen vähentää oppilaalta huomattavasti omaa pohtimista kuinka paljon dataa vain koulusta pystyy keräämään.

Ylöspäin eriytettäessä tehtävään voidaan alkaa ottamaan mukaan esimerkiksi python-kielellä ohjelmointia. Tällöin taulukkolaskentaohjelman sijaan data sijoitettaisiin listoihin ja oppilaille voidaan antaa tehtäväksi laskea keskiarvovektori python koodin avulla. Oppilaille voidaan antaa myös mahdollisuus etsiä itse mielenkiintoisia suurdatajoukkoja ja tietoa mihin niitä voidaan käyttää.

3.3 Tehtävänanto

Kerää lähimmästä kymmenestä koulusta mahdollisimman monta piirrettä taulukko-ohjelmaan. Aseta koulujen nimet ensimmäiselle riville ja piirteet ensimmäiselle sarakkeelle. Täytä tiedot piirteittäin koulun nimen alle. (Jätä tyhjäksi tiedot, joita ei voida määrittää tai ei löydy. Piirteet voivat olla lukuarvoja kuten rakennuksen ikä, oppilasmäärä

ja sisäänpääsyraja, luokittelevia kuten kunnallinen / valtiollinen (esimerkiksi normaalikoulu) / yksityinen koulu, peruskoulu / lukio / ammattikoulu / yliopisto tai molempia esimerkiksi koulu keskustassa / etäisyys keskustasta.)[1]

a. Kuinka monta piirrettä on yhdellä datapisteellä enimmillään?

b. Millainen olisi alueesi keskivertokoulu? Luo tälle oma sarake (käytä luokkien kohdalla yleisintä luokkaa eli moodia ja lukuarvojen kohdalla keskiarvoa).

c. Pohdi mitkä tiedot voisivat kiinnostaa seuraavia tahoja ja miksi: Suomeen muuttavan perheen vanhemmat, yläasteelta valmistuva nuori, stipendejä jakava säätiö.

Kesto: 20–30min opettajajohtoinen johdanto vektoreista sekä tehtävänanto, ja 55–45min tehtävän tekoa luokassa. 2–4 päivää kotitehtävänä.

Opettaja jakaa tehtävänannon ja näyttää omasta koulusta esimerkin taulukkolaskentaohjelmaan. Omasta koulusta kannattaa käydä ainakin yksi lukuarvollinen ja yksi luokitteleva piirre, mutta oppilaiden olisi hyvä pohtia itse datan keräämistä. Mikäli lukion lähellä on huomattavan vähän kouluja, tehtävään voidaan ottaa lähin kaupunki.

3.3.1 Ongelmalähtöinen lähestymistapa

Kesto: 75min opettajan esimerkit vektoreista ja tämän avulla ryhmätyön aloitus, 2–4pv kotitehtävänä.

Oppilaat työskentelevät 3–5 henkilön ryhmissä eivätkä saa suoraan suluilla merkittyä osaa tehtävänannonohjeista tyhjistä tiedoista tai datan muodosta lukuarvona tai luokkina. Nämä tehtävänosat ovat ongelmia joihin oppilaat saavat itse tai opettajan johdattelemana törmätä. Opettajan tehtävä ongelmalähtöisessä menetelmässä onkin nimenomaan saada oppilaat pohtimaan miten erilaisia ja moninaisia piirteitä datasta voidaan löytää.

Määrittäessään keskivertokoulua oppilaille ei myöskään anneta ohjeita käyttää moodia, vaan tämä ongelma pitää osata itse ratkaista. Oppilaille lienee helpointa löytää juuri tämä ”yleisin luokka” ratkaisu, mutta esimerkiksi koulutusasteelle voi teknisesti määrittää lukuarvon, esimerkiksi päiväkoti=1, alakoulu = 2 jne. ja laskea näiden avulla keskiarvon. Tämä lähestymistapa voi vaikuttaa toimivalta tilanteessa,

[1] Tässä tutkielmassa merkitään suluilla tehtävänannon osat, jotka jätetään pois ongelmalähtöisessä opetuksessa.

jossa kouluja on melko tasaisesti kaikilta asteilta, mutta aina tämä lähestymistapa ei toimi. Esimerkiksi jos kunnallinen koulu saa arvon 1, normaalikoulu arvon 2 ja yksityinen koulu arvon 3, keskiarvo ei enää ole toimiva. Oppilaille tämän voi selittää siten, että miten keskivertokoulu voisi olla valtiollinen koulu, jos kunnassa olisi esimerkiksi kolme yksityistä ja kolme kunnallista koulua, eikä yhtään valtiollista. Jos taas luokittelevista piirteistä löytyy kaksi tai useampi yhtä suurta luokkaa, mikä vain näistä kelpaa moodiksi.

3.3.2 Tehtävän purku

Koska oppilailla voi olla erilaisia vastauksia etenkin ongelmalähtöisessä tehtävänannossa, myös purkamiseen voi olla hyvä käyttää hieman enemmän aikaa, esimerkiksi 20min. Oppilaiden tulee palautta Excel-taulukot edellisenä päivänä, jotta opettajalla on aikaa tutustua heidän vastauksiinsa. Oppilaiden kanssa pohditaan, mitä piirteitä he ovat löytäneet ja missä muodossa he ovat datan päättäneet tallentaa. Purussa on hyvä ottaa puheeksi aiemmin mainitut seikat siitä, miksi lukuarvoille kelpaa kesiarvo, mutta luokkapiirteille pitää määrittää moodi, mikäli niitä ei vielä kaikkien kanssa ole käyty. Kun pohditaan datan piirteitä, jotka voisivat kiinnostaa tehtävänannossa lueteltuja tahoja, on syytä myös pohtia, onko datassa mahdollista ottaa huomioon kaikki, mikä kyseisiä henkilöitä voisi kiinnostaa ja miten yleistettävissä nämä tulkinnat olisivat. Esimerkiksi Suomeen muuttavaa perhettä ehkä kiinnostaisi tietää koulun etäisyys keskustasta joka tapauksessa, mutta kaupungin keskusten asuntojen hinnat voivat olla huomattavasti kalliimpia kuin maalaiskunnan keskustassa tai taajamalla. Näin datasta voitaisiin vetää erilaisia johtopäätöksiä riippuen mistä se on kerätty.

Purkamisen lopuksi opettaja esittelee yhden tunnetuista suurdatajoukoista, kuten esimerkiksi Ottawa Real Estate, European Soccer tai Titanic. Ensimmäinen käsittelee Ottawan asuntomarkkinoita ja joukossa on 1255 pistettä, joilla on kullakin 14 piirrettä esimerkiksi asunnon hinta, koordinaatit ja huoneiden määrä (Kaggle, 2020). European Soccer -datajoukko on massiivinen joukko dataa jalkapallopeleistä, pelaajista, näiden ominaisuuksista ja vedonlyönnistä, ja joukko sisältää kymmeniä tuhansia pelejä, kymmenen tuhannen pelaajan ominaisuudet, joukkueiden kokoonpanot koordinaateissa ja kuvaukset pelistä (Kaggle, 2016). Tämä datajoukko on mittasuhteiltaan ja ominaisuuksiltaan massiivinen ja havainnollistaa hyvin suurdatan luonnetta. Titanic data taas käsittelee nimensä mukaisesti Titanic-laivan mukana olleita ihmisiä ja heidän tietojaan (Kaggle, 2016).

4 Tehtävä 2: Lineaarinen regressio

Tämä tehtävä tutustuttaa oppilaan lineaariseen regressioon – malliin, jossa datasta muodostetaan lineaarinen kuvaus, jolla pyritään ennustamaan vanhan datan pohjalta uusia havaintoja. Tästä esimerkkinä voisi olla esimerkiksi myyntiin tulleen asunnon hinnan ennustaminen alueen muiden asuntojen hinnasta. Lineaarinen regressio on hyvin yksinkertainen, mutta tehokas ohjattu koneoppimismenetelmä ja sitä käytetään usein perusosana monimutkaisimmissa neuroverkkoissaa.

4.1 Tehtäväkohtaiset tavoitteet

Oppilas tutustuu siihen, miten lineaarinen regressio, sekä monet muut koneoppimismenetelmät toimivat algoritmisesti, sekä mitä tarkoittavat käsitteet lineaarikuvaus, parametri, virhefunktio, gradienttimenetelmä, ja mikä näiden merkitys koneoppimisessa. Kukin näistä esitellään kaksiulotteisen esimerkin kautta, mutta oppilaille on syytä selventää, että oikeasti datalla on joskus monta sataa piirrettä, kuten tehtäväpaketin edellisessä suurdatatehtävässä on käyty läpi. Käsitteitä ei myöskään tarvitse osata syvällisellä tasolla, vaan oppilaalle riittää muistaa ne esimerkiksi osana vuokaaviota. Olennainen tavoite on osata arvioida koneoppimisen ja lineaarisen regression sopivuutta annettuun ongelmaan; esimerkiksi kykeneekö malli ottamaan huomioon kaiken tarvittavan tai onko ongelma ylipäätään ratkaistavissa lineaarisella regressiolla.

4.2 Materiaalit ja eriyttäminen

Tehtävänantoon liittyy ympäristöhallituksen ylläpitämältä asuntojen.hintatiedot.fi sivustolta haettu taulukko Turun keskustan alueella myydyistä asunnoista sekä näiden ominaisuuksista: huoneiden määrä, neliömäärä, myyntihinta ja rakennusvuosi. Opettaja voi myös käyttää kyseistä sivua etsimään oman kunnan alueelta tai kaupungin alueelta, mutta tällöin on hyvä huomata, että etäisyys kaupungin keskuksestaan on varsin merkittävä piirre. Tehtävänannossa odotetaan neliömäärän ja hinnan olevan suhteellisen lineaarisia, joten opettajan on hyvä tarkistaa itse kerätyn datan osalta, että tämä toteutuu. Mikäli opettajalla on taitoa tehdä datasta oikea moniulotteinen lineaarinen regressio, kaikki piirteet kannattaa tietenkin ottaa mukaan, vaikka dataa täytyykin tällöin hieman muokata manuaalisesti. Koneoppimismallin esitteleminen on suositeltavaa, sillä tehtävä ei itsessään demonstroi aiheen tietoteknistä osaamista tai käytännön työskentelyä, joita molempia suositellaan tekoälykoulutukseen (UNESCO, 2022).

Opettajan tulisi myös pystyä selittämään vakuuttavasti tehtävässä käytetyt termit, jotka löytyvät myös tutkielman alusta, sekä näyttämään miten lineaarinen kuvaus yksinkertaistuu yhdellä tutkittavalla piirteellä ja kahdella parametrillä suoran sovittamiseksi. Tämän lisäksi opettaja voi selittää neliökeskivirheen merkityksen graafisesti suoran sovittamisen kautta, mutta etenkin ongelmalähtöisessä opetustavassa, oppilaiden olisi hyvä päätyä itse tutkimalla virheen minimoimisen merkitykseen.

4.2.1 Eriyttäminen

Tehtävä antaa monipuolisesti mahdollisuuksia eriyttää joko valinnaisen ohjelmoinnin tai tehtävänannon mukaan. Oppilaille on mahdollista teettää esimerkiksi Jupyter notebook, jonka avulla oppilaat voivat tehdä oikean lineaarisen regression esimerkiksi muokkaamalla valmista ohjelmakoodia. Koska tehtävä on hyvin laaja, tällaista ei esitellä osana tätä tutkielmaa. Ylöspäin voidaan eriyttää myös antamalla oppilaiden laskea derivaatta hintojen lineaarikuvauksen neliökeskivirheelle. Tällöin oppilaat joutuvat pohtimaan derivaatan toimintaa summamerkinnän kanssa.

Tehtävän alaspäin eriytettäessä oppilaille voidaan esitellä gradienttimenetelmä ensin opettajan johdolla, sillä tämä on helppo algoritmi, mutta sen päätteleminen pelkästään kaavasta vaatii aikaa ja pohtimista oppilaan tasolta. Opettaja voi myös näyttää oppilaille, miten taulukkolaskentaohjelmalla voidaan laskea neliökeskivirhe, jolloin oppilaiden ei tarvitse päätellä tätä itse.

4.3 Tehtävänanto

Vektorimuotoisesta datajoukosta $X = \{\bar{x}_0, \bar{x}_1, \bar{x}_2, \dots, \bar{x}_m\}$ voidaan valita yksi merkittävä piirre, joka halutaan ennustaa muiden piirteiden avulla. Tätä ennustettavaa arvoa kutsutaan nimellä *leima* ja sitä merkitään yleensä kirjaimella y. Data myös jaetaan harjoitusdataan, validointidataan ja testidataan.

Koneoppimisen yksinkertaisin algoritmi tunnetaan nimellä lineaarinen regressio. Lineaarisessa regressiossa muodostetaan datasta ennustava funktio, joka on muotoa

$$f(\bar{x}_i) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n.$$

Tässä lineaarisessa kuvauksessa datapisteen piirteet, eli siis vektorin komponentit, kerrotaan jollakin parametrilla θ ja summataan yhteen mukaan lukien myös vakioparametri θ_0, joka ei toimi kertoimena millekään datan piirteelle.

a. Kopioi liitteen data asuntojen hinnoista ja ominaisuuksista taulukko-ohjelmaan. Tutki datasta, mikä piirre kuvaa parhaiten asunnon hintaa suoralla. Sovita näihin pisteisiin suora. Mikä on suoran yhtälö?
b. Mikä olisi oman kotisi hinta tämän suoran perusteella? Onko tämä hyvä keino arvioida talon hintaa? Perustele vastauksesi.

Huomaat, että suora ei sovi täysin datapisteisiin. Tätä eroa suoran pisteiden ja datapisteiden välillä kutsutaan virheeksi. Yksi yleisesti käytetyistä virhefunktioista on keskineliövirhe

$E_n = \frac{1}{N}\sum(f(\bar{x}_i) - y_i)^2$, jossa summa käy läpi kaikkien datapisteiden lineaarikuvauksen arvon ja niitä vastaavat leimat.

b. Laske suoran funktiolla saatujen arvojen ja datan leimojen välinen keskineliövirhe taulukko-ohjelmalla.
c. Kuvaile miten ratkaiset, millä kulmakertoimen arvolla virhefunktio saa pienimmän arvonsa? Entä millä vakiotermin arvolla?

Usean muuttujan funktion derivoimista tietylle muuttujalle kutsutaan osittaisderivoinniksi. Derivaatan nollakohdan ratkaiseminen algebrallisesti on kuitenkin usein turhan työlästä, ja algoritmi tehdään joka tapauksessa suuren laskentatehon tietokoneella. Virhefunktion minimi etsitään usein numeerisesti niin sanotulla gradienttimenetelmällä

$$\theta_{i\ uusi} = \theta_i - \alpha\frac{\delta E}{\delta\theta_i},$$

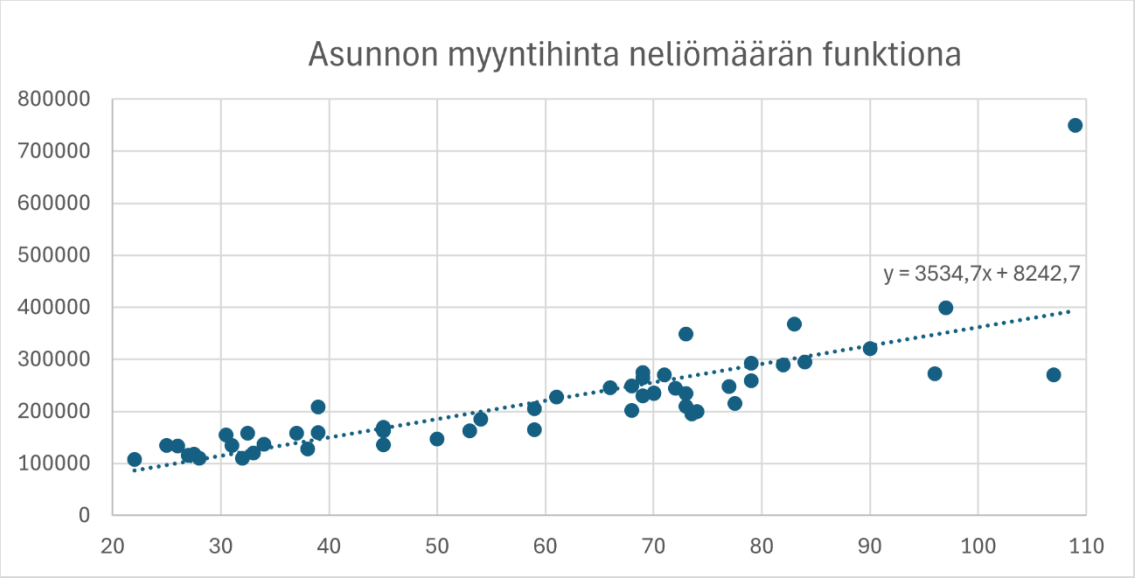
jossa α on jokin vakio, ja $\frac{\delta E}{\delta\theta_i}$ edellä mainittu virhefunktion osittaisderivaatta. Parametrejä päivitetään, kunnes ne eivät enää muutu merkittävästi, jonka jälkeen lineaarikuvauksen toimivuus varmistetaan validointidatalla. Tämän jälkeen malliin tehdään tarvittavat muokkaukset.

d. Piirrä paraabeli $g(x) = x^2 + 1$. Kuinka monta gradienttimenetelmän askelta tarvitset löytääksesi funktion minimin yhden desimaalin tarkkuudella, kun lähtöpiste on $x = 4$ ja $\alpha = \frac{1}{4}$?
e. Entä jos $\alpha = 1$?
f. Piirrä vuokaavio lineaarisesta regressiosta (seuraavilla askelilla: parametrien päivittäminen gradienttimenetelmillä, datan valmisteleminen, lineaarikuvauksen muodostaminen, neliökeskivirheen laskeminen, mallin tarkastaminen validointidatalla ja mallin testaaminen testidatalla, mallin hylkääminen, mallin hyväksyminen, datan kerääminen)

Kesto: 15-30min opetus miten datan piirteistä voidaan muodostaa lineaarinen kuvaus, 45min tehtävän tekoa tunnilla sekä 2-4pv kotitehtävänä.

Kuten edellä mainittu, tehtävä antaa opettajalle paljon valinnanvaraa siihen, kuinka paljon tämä haluaa jättää oppilaiden itse selvitettäväksi. Tehtävän johdanto voidaan vielä avata opettajajohtoisesti, etenkin jos oppilailla on vaikeuksia hahmottaa, miten parametrit, piirteet ja vektorit toimivat yhdessä ja

miksi kutakin voidaan käsitellä muuttujana eri tehtävän vaiheessa. Suoran sovittaminen taulukkolaskentaohjelmalla, kuten kuvassa 2 on myös syytä kerrata, mikäli tämä ei ole oppilaille selvää. Oppilaille on myös syytä selittää, miten vektorit $\bar{x}_i$ indeksoidaan välillä $[0, \ldots, \mathrm{m}]$ ja näiden piirteet x_j välillä $[0, \ldots, \mathrm{n}]$. Oppilaalle voi olla sekavaa, milloin puhutaan parametreista, milloin datan vektoreista ja milloin piirteistä, jolloin kaavojen ymmärtäminen on huomattavasti vaikeampaa, kuin sen tarvitsisi. Gradienttimenetelmän simuloinnissa opettaja voi myös näyttää yhteisesti, miten algoritmia käytetään normaalin funktion kanssa. Helpoin keino laskea neliökeskivirhe, mikäli taulukkolaskentaohjelmassa ei ole tälle valmista kaavaa, on syöttää taulukko-ohjelmalla sovitetun suoran funktioon datapisteet, tämän jälkeen laskea neliövirhe jokaiselle pisteelle erikseen, summata nämä yhteen ja jakaa summa pisteiden määrällä taulukon 2 mukaisesti. Kuva on katkaistu tilan säästämiseksi.



Kuva 2: esimerkki taulukko-ohjelman sovituksesta.

Taulukko 2: Esimerkki neliökeskivirheen laskemisesta

Neliöt [m²]	Myyntihinta [€]	Ennustettu hinta (3534,7*Neliöt+8242,7)	Neliövirhe (Myyntihinta-Ennustettu hinta)^2	Neliökeskivirhe: (sum(neliövirhe)/määrä)
28	110000	107214,3	7760124,49	3900308024
39	159000	146096	166513216	
33	120000	124887,8	23890588,84	
25	135000	96610,2	1473776744	
31	134000	117818,4	261844178,6	
32	110000	121353,1	128892879,6	
26	133000	100144,9	1079457596	

4.3.1 Ongelmalähtöinen lähestymistapa

Kesto: 60min tehtävän tekoa tunnilla sekä 2-4pv kotitehtävänä.

Ongelmalähtöisessä opetuksessa tehtävä voidaan jakaa oppilaille jopa suoraan, jolloin ryhmille jää enemmän aikaa pohtia tehtäviä opettajan valvonnassa tunnilla. Huomiot indekseistä, parametreistä ja muuttujista mahdollisina hämmennyksen aiheina pätevät samoin kuin normaalin opetuksen muodossa. Gradienttimenetelmän simulointi tuottaa myös vaikeuksia, mikäli oppilaat eivät ymmärrä, että tehtävässä tulee tehdä algoritmi muuttujan x suhteen, eikä enää vakioiden. Tehtävässä tämä on jätetty tarkoituksella tekijän ongelmaksi, mutta oppilaiden ei tulisi jäädä tähän jumiin siitä huolimatta. Ryhmät saattavat myös koittaa jakaa tehtävän osat keskenään, mutta tämä saattaa olla tämän tehtävän tapauksessa huono lähestymistapa, sillä tehtävän osat rakentuvat edellisten päälle, joten edellisiä kohtia pitää ainakin pohtia yhdessä. Opettajan harkinnan varaan jää kerrotaanko tämä, vai annetaanko oppilaiden mahdollisuus myös pohtia töiden jakamista.

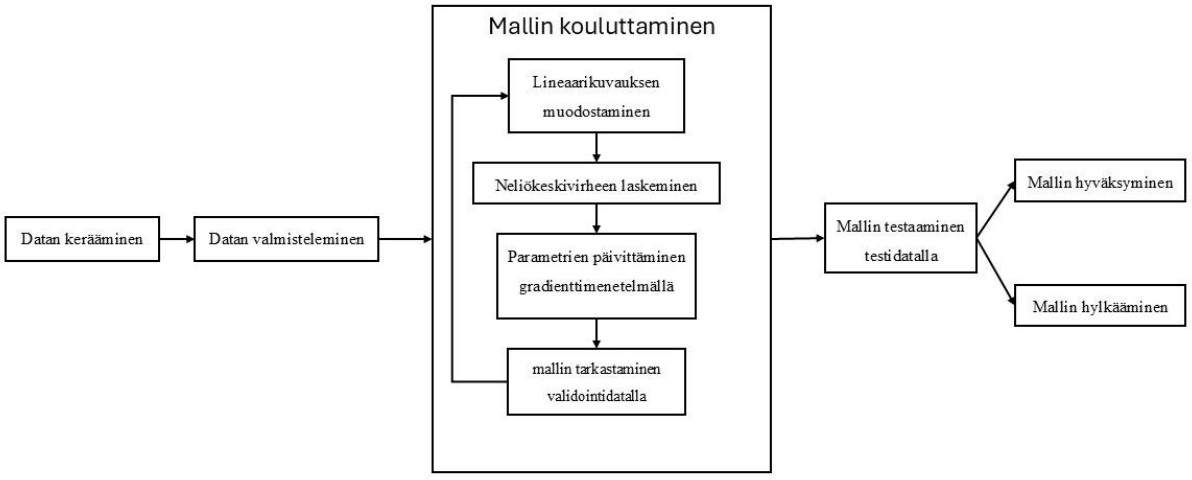
4.3.2 Tehtävän purku

Oppilailta kerätään ennen purkutuntia vastaukset tehtäviin. Taulukko voidaan arvioida suoraan sen mukaan ovatko oppilaan antamat arvot oikein vai väärin, ja että vastauksesta käy ilmi, että suoralla hinnan arviointi ei selvästi ole riittävä asunnon hinnan arvioija. Oppilaat saattavat perustella tämän esimerkiksi sillä, että asunnon hintaan vaikuttaa huomattavasti enemmän tekijöitä kuin neliömäärä ja taulukossa olevat oudokit todistavat näyttävät tämän riittävästi. Opettaja voi näyttää tässä vaiheessa demonstraation monen piirteen lineaarisesta regressiosta, joka ennustaa asuntojen arvoja huomattavan hyvin ja jatkaa keskustelua, että olisiko tämä nyt hyvä keino päättää asuntojen hinnat. Tämä kysymys voi jo jakaa mielipiteitä ja opettaja voi johdatella keskustelun koulutukseen käytetyn datan puolueettomuuteen esimerkiksi kysymyksillä: ”Osaako tämäkään malli ottaa huomioon kaiken ihmisten asumiseen liittyvän?” tai ”Osaisiko tämä malli huomioida, että jokin kaupungin alue saattaa olla syrjityn vähemmistön asuttamaa?”.

Pohdintaa voidaan tehdä myös ilman opettajan esimerkkiä regressiomallista, sillä tarkoituksena on herättää pohdintaa siitä, miten tekoälymallit pohjautuvat aina johonkin edeltävään dataan ja niiden tuloksia tulee tulkita dataan perustuvina ennusteina, eikä absoluuttisena totuutena. Opettaja voi esimerkiksi esitellä sivulta löytyvää dataa, mutta laajemmin koko Suomen alueelta. Tällöin on nähtävissä, miten vajavaista data on tietyille piirteille ja miten paljon heittelyä datan määrässä ja asuntojen tyypeissä nähdään esimerkiksi kaupunkien ja maaseutukuntien välillä. Opettajan johdolla voidaan pohtia, miten datasta jouduttaisiin todennäköisesti kouluttamaan monta erillistä mallia eri asuntotyypeille ja eri kaupungeille.

Gradienttimenetelmän osalta oppilaiden kanssa kannattaa kiinnittää huomiota, miten vakion α valitseminen vaikuttaa algoritmin toimintaan ja miten huonosti valittu vakio voi saada algoritmin suorituksen kestämään huomattavan kauan tai jopa estää suppenemisen. Oppilaat voi myös johdatella neliökeskivirheen konveksisuuteen esimerkiksi kysymällä mitä yhteistä huomataan paraabelin yhtälöllä ja neliökeskivirheellä. Tähän voi auttaa myös unohtamalla hetkeksi summamerkin ja sen jälkeen huomauttamalla, että toisen asteen termien summaaminen ei muuta sitä seikkaa, että kyseessä on paraabeli ja näin ollen konveksifunktio.

Vuokaavion purkamisen kanssa on syytä painottaa datan käsittelyn merkitystä, mikä liittyy myös edeltävään keskusteluun datan merkityksestä tekoälymallien koulutuksessa. Kuvassa 3 on esitelty miltä vuokaavio voi näyttää, mutta yhtä oikeaa muotoilua tälle ei ole. Oppilaat saattavat myös kyseenalaistaa validointidatan ja testidatan erottelun, mutta tämä voidaan selittää sillä, että validointidatan käytön jälkeen mallia voidaan vielä muokata, kun taas testidatan jälkeen malli on joko hyväksyttävä tai hylättävä, jolloin sen rakentaminen pitää aloittaa täysin alusta.



Kuva 3: esimerkki vuokaaviosta.

5 Tehtävä 3: Luokittelu

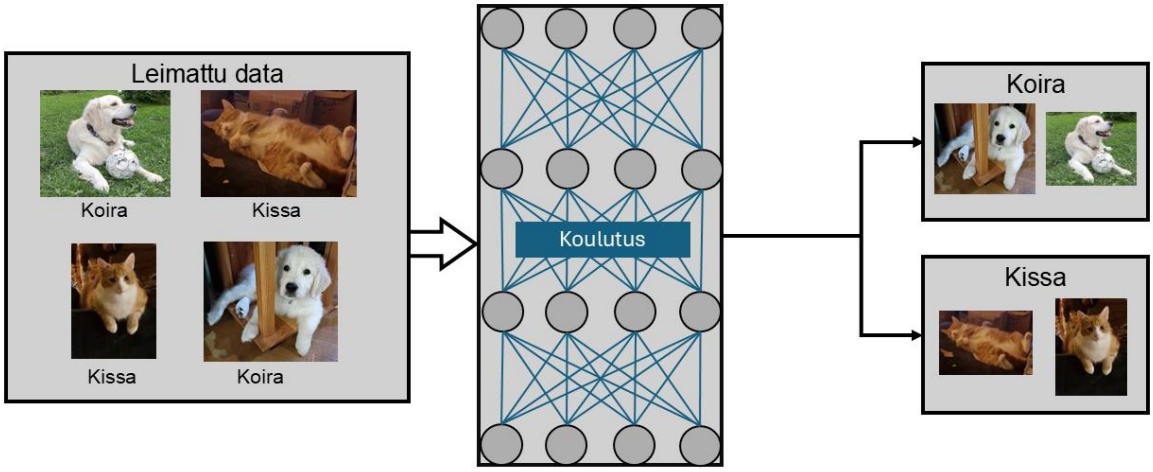
Luokittelutehtävässä lineaarisesta regressiosta siirrytään logistiseen regressioon ja sen käyttöön luokitteluongelmassa. Luokittelu on varsin monikäyttöinen ja paljon käytetty koneoppimismenetelmä myös monimutkaisemmissa koneoppimismalleissa, sillä se mahdollistaa paitsi ennustamisen aiemmasta datasta, myös tiedon jäsentelyn monimutkaisemmissa koneoppimismalleissa. Luokittelumallin perusajatus on nimensä mukaan jakaa data piirteiden perusteella edeltä määriteltyihin luokkiin. Tämä voidaan tehdä joko jonkinnäköisellä porrasfunktiolla, jossa jokainen porras tai väli edustaa yhtä luokkaa, mutta yksinkertaisempaa on tarkastella niin sanottua ’’yksi vastaan muut’’ mallia, jossa tarkastellaan yksi kerrallaan pisteen todennäköisyyttä kuulua yhteen luokista, minkä jälkeen luokaksi valitaan todennäköisin.

5.1 Tehtäväkohtaiset tavoitteet

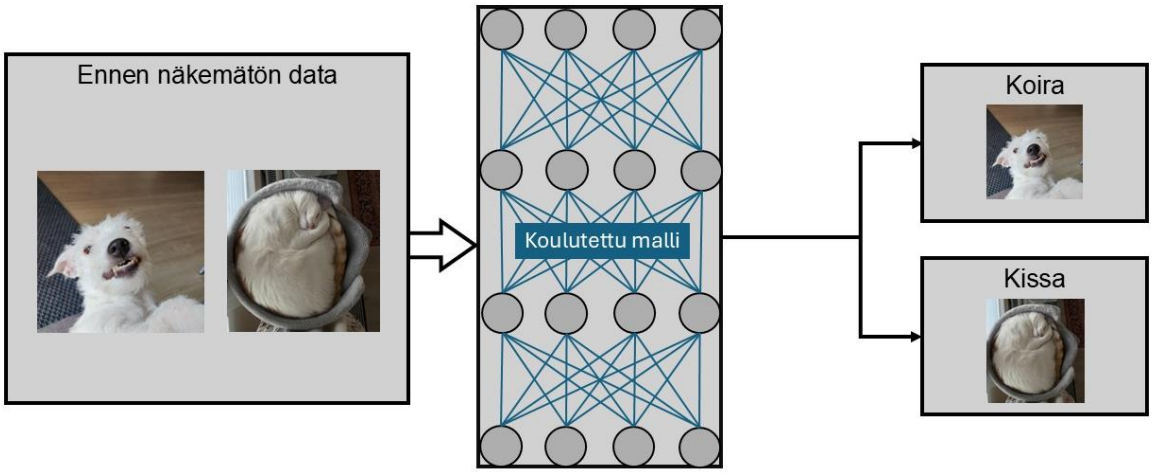
Tehtävä jatkaa samaa teemaa kuin edellinen tehtävä, sillä logistinen regressio pohjautuu lineaariseen regressioon ja monet samat perusalgoritmit toimivat pääpiirteittäin. Päätavoitteet oppilaan osaamiselle ovat logistisen funktion muoto, logaritmivirheen tarkoitus, soveltaminen kaksiulotteisessa tapauksessa sekä luokittelualgoritmin kuvaileminen. Luokittelun yhteys lineaariseen regressioon tulee tehtävässä myös ilmi, mikä edistää oppilaan ymmärrystä lineaarisesta regressiosta osana muita koneoppimismalleja. Tehtävä on myös viimeinen tehtäväpaketissa käsiteltävä ohjatun koneoppimisen menetelmä, ja vaikka ohjattu ja ohjaamaton koneoppiminen termeinä esitellään vasta seuraavassa tehtävässä, näiden eron sisäistämiseksi oppilaan tulee tiedostaa, että luokittelussa mallit edelleen koulutetaan ennalta määritettyjen leimojen pohjalta. Tehtävä siis toimii paitsi tärkeän algoritmin opettamiseen, myös siltana koneoppimisen eri osa-alueiden välillä.

5.2 Materiaalit ja eriyttäminen

Opettaja voi esitellä luokitteluongelman tunnilla esimerkiksi seuraavanlaisesti: Luokitteluongelma on toinen hyvin olennainen koneoppimisen osa-alue. Luokittelussa mallille esitellään leimattuja datapisteitä, joista malli oppii jakamaan piirteiden perusteella uuden datan luokkiin. Esimerkkinä voisi olla koirien ja kissojen tunnistaminen, mikä on varsin yksinkertainen tehtävä ihmiselle, mutta tietokoneelle huomattavasti vaikeampi. Esim. CAPTCHA tunnistautumiset käyttävät tätä ominaisuutta varmistamaan, että profiilin luo oikea ihminen. Koneoppimisella on kuitenkin mahdollista opettaa myös tietokoneohjelma tunnistamaan kuvista haluttuja ominaisuuksia. Kuvissa 4 ja 5 on esitelty perusteet, miten luokittelumalli koulutetaan ensin valmiiksi leimatulla datalla, minkä jälkeen mallin tulisi osata luokitella ennen näkemättömiä kuvia.



Kuva 4: Luokittelumallin kouluttaminen.



Kuva 5: Koulutettu luokittelumalli.

Tehtävässä käytetään Wisconsinista kerättyä rintasyöpädataa, mutta data on rajattu neljään piirteeseen ja noin kahteen sataan pisteeseen, jotta sen käsittely olisi tehtävän kannalta mielekästä (Wolberg, Mangasarian, Street, & Street, 1993). Data on kuitenkin hyvä esitellä oppilaille esimerkiksi Kagglen kautta.

5.2.1 Eriyttäminen

Tehtävän data on mahdollista hakea internetistä itse ja jopa avata taulukkolaskentaohjelmalla, mutta dataa voi olla epäkäytännöllistä muokata yhtään heikommalla tietokoneella. Mikäli oppilaiden tietokoneiden laskentatehot riittävät, heille voi antaa tehtäväksi hakea data itse, ja riippuen oppilaiden

tasosta, muokata se oikeaan muotoon joko Pythonilla tai taulukkolaskentaohjelmalla. Pythonilla datasta voidaan ensin karsia ylimääräiset piirteet ja tämän jälkeen valita datasta tietty otos, minkä jälkeen tämä otos voidaan siirtää taulukkolaskentaohjelmaan. Taulukkolaskentaohjelmalla data pitää erotella pilkkujen avulla sarakkeille, mutta tällöin on tärkeää huomata, että lukuarvot saattavat olla väärässä muodossa, jolloin ne kannattaa ensin tallentaa tekstinä, muuttaa desimaalierottimet pilkuiksi ja muuttaa vasta sen jälkeen luvuiksi. Mikään ei myöskään estä periaatteessa käyttämästä koko dataotosta, sillä sen perusteella laskettavat tehtävät ovat melko helppoja määrästä huolimatta.

Alaspäin erityttäessä opettaja voi yksinkertaisesti antaa enemmän ohjeita esimerkiksi siitä, miten todennäköisyydet lasketaan, logistinen funktio piirretään ja siitä, miltä logistinen funktio näyttää, kun siihen upotetaan lineaarinen kuvaus, mikä on tehtävässä jätetty oppilaiden selvitettäväksi. Opettaja voi myös muotoilla taulukoidun datan esimerkiksi kasvaimen koon mukaan tai vaikka jakaa sen valmiiksi neljään kokoluokkaan, mikäli datan käsittelyyn kuluva aika halutaan pitää minimissään. Oppilaille voi myös antaa vinkin, että yksi keino laskea helposti pahanlaatuisten kasvaimien määrä, on korvata hyvä- ja huonolaatuisen merkit luvuilla 0 ja 1. Tämä tehtävä antaa myös mainion mahdollisuuden demonstroida logistista regressiota koodaustehtävien kautta, mutta tämä jätetään tämän tutkielman puitteissa käymättä.

5.3 Tehtävänanto

Luokittelualgoritmissa pyritään jakamaan ennen näkemättömiä datapisteitä ennalta määrättyihin luokkiin. Yleisesti käytetty binäärinen luokittelun keino on tehdä ensin datalle lineaarinen kuvaus $f(\bar{x}_i)$ ja syöttää tämä lineaarinen regressio niin sanottuun logistiseen funktioon

$$L\big(f(\bar{x}_i)\big) = \frac{1}{1 + e^{f(\bar{x}_i)}}.$$

Logistisessa regressiossa ei käytetä parametrien päivittämiseen mitattavaa virhettä, vaan algoritmia rankaistaan tappiofunktiolla.

a. Tutki Geogebran avulla kahdella parametrilla, miksi logistinen funktio ja lineaarikuvaus sopivat erityisen hyvin kahden luokan luokitteluun, mainitse kolme syytä.
b. Yleisenä tappiofunktiona logistiselle funktiolle käytetään niin sanottua logaritmitappiota, joka saa yhdelle pisteelle muodon

$$E_{log} = y_i \cdot ln\big(L(\bar{x}_i)\big) - (1 - y_i) \cdot ln\big(1 - L(\bar{x}_i)\big).$$

Tutki Geogebralla miksi.
(vinkki: mitä virheelle tapahtuu, kun logistisen regression ennuste on oikeassa/väärässä?)

Liitteenä on Wisconsinissa kerätty taulukko rintasyöpäkasvainten koosta ja laadusta. Alkuperäisessä datassa on 32 piirrettä ja 569 datapistettä, mutta data on tiivistetty tässä kahteen sataan pisteeseen ja kolmeen piirteeseen sekä tunnisteeseen. Yritetään ennustaa kasvainten koosta, onko se hyvän- (H) vai pahanlaatuinen (P).

c. Kopioi edellä mainittu taulukkolaskentaohjelmaan. Valitse käytettäväksi piirteeksi joko kasvaimen pinta-ala tai säde. Perustele valintasi.

d. Jaa data neljään eri kokoluokkaan valitsemasi piirteen mukaan ja laske millä todennäköisyydellä kunkin kokoluokan kasvain on pahanlaatuinen.

e. Sovita Geogebralla näihin todennäköisyyksiin logistinen käyrä (käytä x-akselina valitsemasi piirteen kokoluokan keskiarvoa). Mitkä ovat parametrit?

Luokitteluongelmassa voi olla myös enemmän kuin kaksi luokkaa, mutta tämä ei vaikuta merkittävästi käytettävään matematiikkaan. Selvitä mitä tarkoittaa ns. ”1 vs. all” luokittelu. (Logistinen regressio voidaan vain tehdä jokaiselle luokalle erikseen ja valita luokka, joka saa suurimman todennäköisyyden.)

f. Piirrä vuokaavio neljän luokan luokittelualgoritmista logistisella regressiolla

Kesto: 15-30min opetus luokitteluongelmasta ja logistisesta funktiosta, 45min tehtäväntekoa tunnilla ja 2-4pv kotitehtävänä.

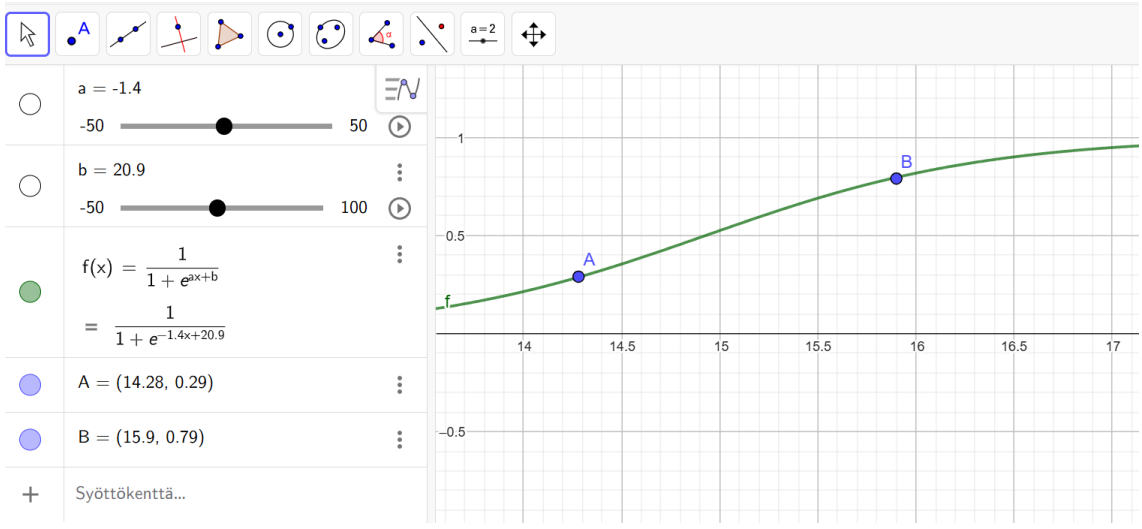
Tehtävän ratkaisun eteneminen vaihtelee huomattavasti riippuen, kuinka paljon oppilaille jätettiin työstettäväksi datan käsittelyä. Mikäli oppilaat lataavat itse datan ja muokkaavat sen käytettävään muotoon, tunnin opetusosuus kannattaa käyttää nimenomaan tarkistamaan, että kaikki saavat datan ladattua ja oikeaan muotoon. CSV-tiedoston avaaminen taulukkolaskentaohjelmalla voidaan jopa näyttää opettajan johdolla. Tehtävän onnistumisen kannalta on kuitenkin kriittistä, että oppilaat saavat datan ladattua järkevässä muodossa koneilleen, joten tähän kannattaa käyttää aikaa. Opettajan tulee myös pitää huoli, että kaikki tajuavat logistisen funktion ja lineaarikuvauksen yhdistämisen, piirtämisen sekä tulkitsemisen. Hyvä lähestymistapa on kirjoittaa funktio Geogebraan muotoon $\frac{1}{1+e^{ax+b}}$, jolloin Geogebra luo parametreille suoraan liukusäätimet. Kun parametrejä selvitetään, oppilaille voidaan huomauttaa, että molempien parametrien puolittaminen saa käyrän loivenemaan, kun taas yksittäin muutettuna ne siirtävät käyrää.

Kun oppilaat alkavat tehdä päätelmiä datasta, heitä voi johdatella valitsemaan oikea piirre sen perusteella kumpi näyttää päällisin puolin jakavan tulokset mahdollisimman hyvin. Tämä ei ole täysin selvää datasta, eikä kaikkien ryhmien tarvitse tulla samaan lopputulokseen, vaan olennaista on nimenomaan datan laadun pohtiminen. Oppilaat voivat perustella, että näillä ei ole merkittävästi väliä, sillä säde ja pinta-ala korreloivat vahvasti toistensa kanssa, mikä on hyvä huomio, mutta oppilaille voi

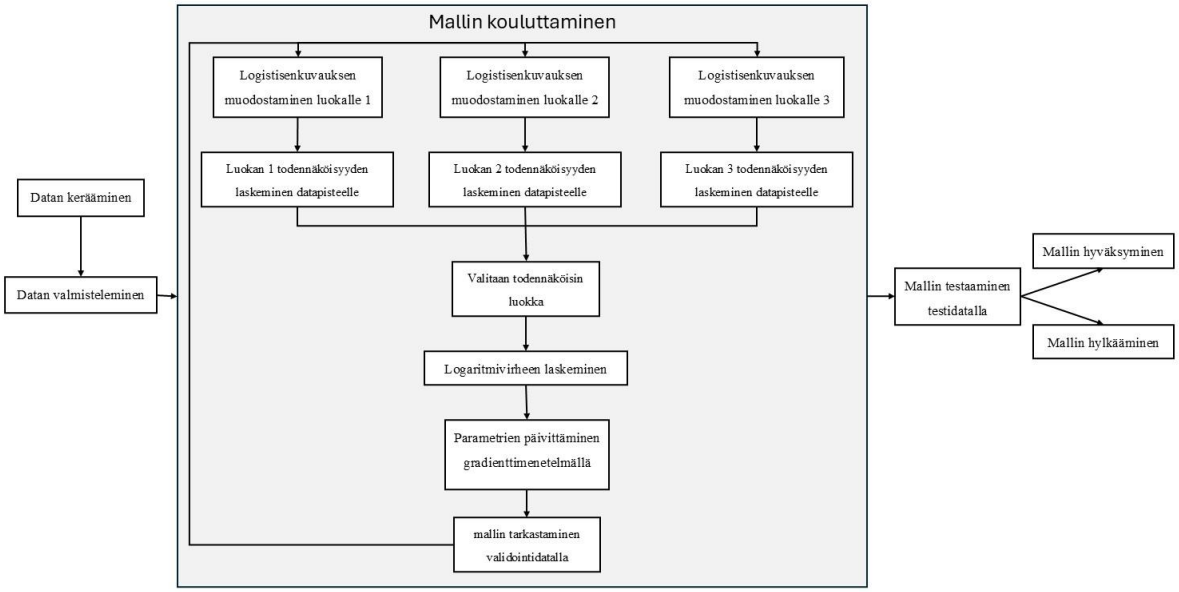
muistuttaa, että pahanlaatuiset kasvaimet ovat usein epäsäännöllisiä, kun taas hyvänlaatuiset pyöreitä, jolloin keskiarvoinen säde ei suoraan kerro pinta-alaa. Todennäköisyyksiä laskiessa oppilaille voidaan myös sanoa, että luokkien ei tarvitse olla samankokoisia ja että kaksi luokkaa voivat olla esimerkiksi lähes aina hyvänlaatuisia ja lähes aina pahanlaatuisia. Alla olevissa kuvissa 7–9 on esitelty esimerkkivastauksia, kun kasvainten kokoluokat on laskettu väliltä 13–15 sekä 15–17.

F	G	H	I	J	K	L	M	N
diagnoosi	**säde (ka)**				**diagnoosi**	**säde (ka)**		
0	13,21				0	15,10		
0	13,21	n:	52		1	15,12	n:	43
0	13,64	Pahanlaatuiset:	15,00		1	15,13	Pahanlaatuiset:	34
0	13,65	P:	0,29		0	15,19	P:	=N4/N3
0	13,66	ka:	14,27640566		1	15,22	ka:	15,9021219
0	13,66				0	15,27		
0	13,68				1	15,28		
0	13,69				1	15,30		
1	13,77				1	15,32		
0	13,77				1	15,34		
0	13,78				1	15,37		
1	13,81				1	15,46		
0	13,85				1	15,46		

Kuva 7: esimerkkilaskut kasvainten säteen kokoluokista 13–15 ja 15–17.



Kuva 8: Esimerkkisovitus logistiselle käyrälle lasketuista keskiarvoista.



Kuva 9: Esimerkki logistisen regression vuokaaviosta

5.3.1 Ongelmalähtöinen lähestymistapa

Kesto: n. 15min opetus luokitteluongelmasta ja logistisesta funktiosta, 60min tehtävän tekoa tunnilla sekä 2-4pv kotitehtävänä.

Tämän tehtävän tehtävänannot eivät muutu merkittävästi ongelmalähtöisessä lähestymistavassa. Ainoat erot ovat suluissa merkityt lisäohjeet, joita ongelmalähtöisessä lähestymistavassa ei anneta. Tällöin oppilaille korostuu tiedon hakeminen oma-aloitteisesti ja he saattavat törmätä ongelmiin esimerkiksi suuruusluokkien määrittämisessä. Ongelmalähtöisestä lähestymistavasta huolimatta, oppilaat eivät saisi jumittua datan kopioimisen tai noutamiseen vaikeuteen, vaan opettajan tulee antaa tässä riittävästi tukea. Oppilailta ei voi odottaa esimerkiksi CSV-tiedoston avaamista ja noutamista itse, eikä tämä osuus ole keskeinen tehtäväpaketin tavoitteisiin. Oppilaat saattavat saada myös huomattavan erilaisia vastauksia parametreille, riippuen miten he ovat määrittäneet kokoluokat, jolloin opettajan kannattaa pyytää heiltä myös kuvakaappaukset kuvaajista.

5.3.2 Tehtävän purku

Tehtävän ensimmäisiä kohtia purettaessa oppilaiden vastaukset voivat vaihdella avoimen tehtävänannon takia. Oppilaille jäädä mieleen logistisen funktion käytöstä, että sen arvojoukko on välillä [0,1], minkä takia se kuvastaa hyvin todennäköisyyksiä, että funktio on monotoninen ja mahdollisuus muuttua joko jyrkemmin tai loivemmin riippuen, miten parametrit valitaan. Kahdella luokalla riittää usein tarkastella vain toista, eli esimerkiksi mikäli kuvassa ei ole koiraa, se on automaattisesti kissa.

Logaritmitappiosta oppilaiden tulisi taas huomata, että virhe on pieni, mikäli ennustettu luokka on lähellä leimaa ja suuri, mikäli ennustettu luokka on väärä.

Rintasyöpäkasvainten kanssa on syytä ottaa puheeksi, miten oikeastaan valinta pinta-alan ja säteen välillä on tosiaan vaikea ja miten nämä suureet selkeästi korreloivat toistensa kanssa, mutta eivät selvästi anna täysin samaa tietoa. Opettaja voi halutessaan näyttää demonstraation miten alkuperäisen datan piirteet korreloivat toistensa kanssa ja mitkä niistä ovat merkittävimpiä piirteitä, mikäli opettajalla on taitoa tehdä tästä riittävän selkeät kuvaajat. Oppilaiden kanssa on myös syytä keskustella käytetyn teknologian etiikasta, sillä puhe on kuitenkin hengenvaarallisen syövän diagnosoimisesta. Voidaan myös miettiä, onko vaarallisempaa diagnosoida hyvänlaatuinen kasvain pahanlaatuiseksi vai toisin päin, missä määrin edes hyvä koneoppimismalli on sovellettavissa lääketieteessä ja kenellä on vastuu, jos virhe tapahtuu. Vuokaavio voi myös tuottaa hyvin erilaisia vastauksia, sillä tällä kertaa vaiheita ei anneta valmiiksi ja oppilaat saattavat esimerkiksi sekoittaa missä vaiheessa virhe lasketaan ja parametrit päivitetään. Esimerkkivastauksessa kuvassa 9 kaikki kolme luokkaa on otettu omiin vaiheisiinsa, jotta olisi selkeämpää, että ne lasketaan erikseen, mutta periaatteessa tämä vaihe voidaan asettaa yhteen jonoon.

6 Tehtävä 4: ryvästys

Tässä tehtävässä oppilas tutustuu ohjaamattomaan koneoppimiseen ja ryvästykseen. Ryvästyksellä tarkoitetaan datan luokittelua ilman, että luokkia on määritelty ennalta. Tätä kutsutaan niin sanotusti ohjaamattomaksi koneoppimiseksi. Ohjaamattomassa koneoppimisessa datasta pyritään löytämään säännönmukaisuuksia sen sijaan, että datasta pyrittäisiin ennustamaan sääntöjen kautta ennalta määrättyjä ominaisuuksia, joita voitaisiin verrata leimoihin. Ohjaamattomalla koneoppimisella on ollut suuri merkitys suurdatan käsittelyssä, sillä usein selkeää yksikäsitteistä oikeaa vastausta tarkasteltuun ongelmaan ei ole. Näin on esimerkiksi seuraavan kysymyksen kanssa: ”Minkä kappaleen musiikkisovelluksen käyttäjä haluaisi seuraavaksi kuunnella?”

6.1 Tehtäväkohtaiset tavoitteet

Oppilaan tulee ymmärtää konseptuaalinen ero ohjatun ja ohjaamattoman koneoppimisen välillä, sekä tietää kolme yksinkertaisinta klustrointimenetelmää: hierarkkinen, keskiarvomenetelmä- ja jakaumaklusterointi. Matemaattisesti näitä menetelmiä ei tarvitse osata, mutta oppilaalle pitäisi jäädä graafinen intuitio, miten nämä menetelmät toimivat kaksiulotteisessa ympäristössä. Ohjaamattoman koneoppimisen yhteys mainontaan, sosiaalisen mediaan ja viihteeseen on myös keskeinen teema, sillä useat alustat tai palvelut käyttävät hyvin todennäköisesti juuri ohjaamattoman koneoppimisen algoritmeja kohdentamaan mainontaa ja sisältöä käyttäjilleen. Tämä tarkoittaa, että käyttäjällä itsellään on pienempi rooli oman vapaa-aikansa käytössä eikä käyttäjän kulutusta välttämättä ohjaa edes ihminen, mihin liittyy useita eettisiä pulmia vastuusta addiktioihin. Opettajan vastuulle jääkin saada oppilaat kyseenalaistamaan kuinka suuri osa heidän elämästään on algoritmien vaikutuksen alaisena.

6.2 Materiaalit ja eriyttäminen

Tehtävää on syytä pohjustaa opetuksella ohjaamattomasta koneoppimisesta. Tätä varten voidaan esittää esimerkiksi ongelma: musiikin kuunteluohjelmasta saadaan paljon dataa käyttäjältä ja tästä datasta haluttaisiin ennustaa käyttäjälle suosituksia. Datalle voidaan määrittää vaikka kuinka paljon piirteitä kuten genre, artisti, kappaleen kesto, onko kappale lisätty soittolistaan, kuinka monta kertaa kappale on kuunneltu, kenen kanssa artisti on tehnyt yhteistyötä ja niin edelleen, mutta tämä data päivittyy jatkuvasti, eikä siitä oikein ole löydettävissä yhtä piirrettä, joka kertoo, pitikö käyttäjä valitusta kappaleesta vai ei. Dataan ei siis ainakaan helposti voida soveltaa aiemmin tehtävissä opittuja malleja.

Tällöin datasta yritetään etsiä moniulotteisia ryppäitä, ja ennen kaikkea sääntöjä, jotka kertovat meille mistä nämä ryppäät löytyvät. Jos esimerkiksi saadaan selville, että musiikki, jota käyttäjä kuuntelee, löytyy kolmesta eri ryppäästä, voidaan hänelle esittää kappaleita, jotka löytyvät näistä ryppäistä, mutta joita hän ei ole kuunnellut. Hänelle voidaan myös esitellä kappaleita neljännestä

joukosta, jota hän ei ehkä ole kuunnellut, mutta joka on yleisesti pidetty niiden joukossa, joiden musiikkimaku kuuluu näihin kolmeen ryppääseen. Tämän jälkeen algoritmit päivittyvät sen perusteella, jatkoiko käyttäjä tämän uuden musiikin kuuntelemista, jolloin hän luo uusia datapisteitä uuteen ryppääseen.

Tehtävään kuuluu myös Suomen kymmenen suurimman kaupungin koordinaattien etsiminen ja tehtävä on rakennettu vuonna 2024 tilastokeskuksen tiedon mukaan kymmenen suurimman kaupungin pohjalta sekä Wikipediasta kerätyistä koordinaateista (Kuntaliitto, 2025). Nämä tiedot on taulukoitu taulukkoon 2, ja tämä taulukko voidaan antaa sellaisenaan oppilaille, mikäli tiedot ovat muuttuneet tai opettajalla ei ole aikaa tarkastaa toimiiko tehtävä ajankohtaisilla tiedoilla.

Taulukko 2: koordinaatit suomen kymmeneen suurimpaan kaupunkiin

Kaupunki	Koordinaatit
Helsinki	60.166640739°N, 24.943536799°E
Espoo	60.206376371°N, 24.656728549°E
Tampere	61.497742570°N, 23.761290078°E
Vantaa	60.298133721°N, 25.006641332°E
Oulu	65.013784817°N, 25.472099070°E
Turku	60.451690351°N, 22.266866666°E
Jyväskylä	62.241677684°N, 25.749498121°E
Kuopio	62.892982923°N, 27.688934744°E
Lahti	60.980380564°N, 25.654987962°E
Pori	61.483726303°N, 21.795900114°E

Opettaja voi myös esitellä hierarkkisen ryvästämisen, jakaumaryvästämisen sekä k-keskiarvoryvästämisen pääpiirteisen idean nimeämättä näitä menetelmiä. Tämä voidaan tehdä esimerkiksi piirtämällä janalle tai tasolle tehtävän kuvien mukaiset pisteet ja näyttäen miten pisteitä voidaan lähteä yhdistämään isommiksi ja isommiksi joukoiksi valitsemalla aina kaksi lähintä joukkoa (hierarkkinen), pisteille voidaan sijoittaa kolme todennäköisyysjakaumat siten, että todennäköisyys kuulua näihin jakaumiin on mahdollisimman suuri (jakauma) tai pisteiden avulla voidaan etsiä keskiarvot siten, että pisteet kuuluvat niitä lähimmän keskiarvon määrittämään ryppääseen (k-keskiarvo). Nämä esimerkit voidaan näyttää resurssien puitteissa taululla, tai hienostuneemmin animaatioilla.

6.2.1 Eriyttäminen

Helpoin keino eriyttää tehtävää alaspäin on esitellä ensin ryvästysmenetelmät, jolloin oppilaan harteille ei jää etsiä itse tietoja, miten ne toimivat. Tämä on jopa suotavaa, sillä suurin osa suomenkielisestä materiaalista koneoppimiseen liittyen ei ole suunnattu lukio-opiskelijalle. Opettajan harkinnan varaan jää kuitenkin, kuinka laajasti hän haluaa aiheesta opettaa ennen tehtävän jakamista, sillä tehtävän kohdat a-c vaikeus riippuu huomattavastisiitä, kuinka paljon oppilaalle annetaan

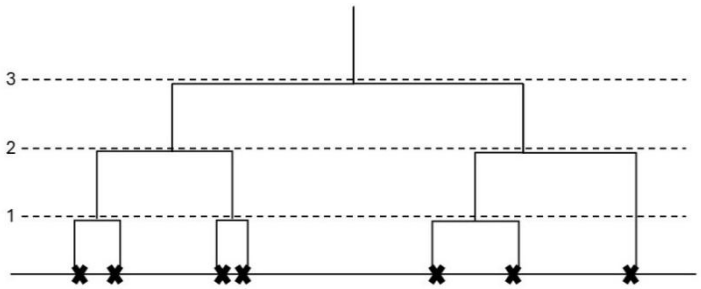
taustatietoa. Oppilaille voi myös kerrata miten lasketaan keskiarvo koordinaateista, mikäli tämä tuottaa vaikeuksia ja oppilaille voi antaa vinkiksi käyttää taulukkolaskentaohjelmaa laskemaan koordinaatteja.

Ylöspäin eriyttäessä oppilaat opettelevat tehtävän aiheen tai etsivät tarvittavat tiedot Suomen kaupungeista itsenäisesti. Edelleen on hyvä muistaa, että ohjaamattomasta koneoppimisesta tiedon etsiminen voi olla haastavaa lukio-opiskelijalle, eikä oppilailta toi odottaa täysin itsenäistä oppimista tässä tehtävässä. Yksi keino päästä yli tästä ongelmasta on tehdä oppilaille valmiit opetusmateriaalit, jotka antavat ryhmän tarpeen mukaan opastusta. Oppilaat voivat myös etsiä itse tiedot tehtävän c-kohtaan, jolloin opettajan tulee tarkastaa itse ajankohtaisten tietojen saatavuus ja k-keskiarvomenetelmän toimivuus.

6.3 Tehtävänanto

Ryvästykseksi kutsutaan koneoppimista, jossa leimaamattomasta datasta etsitään säännönmukaisuuksia eli ryppäitä. Tämä teknologia on usein esimerkiksi viihdesovellusten suositusalgoritmien takana. Tarkastellaan kolmea ryvästysmenetelmää: hierarkkinen-, jakauma- ja k-keskiarvoryvästys.

a. Alla on kuva yhdestä edellä mainitusta ryvästysmenetelmästä. Nimeä se ja tee siitä vuokaavio.



b. Piirrä kolme normaalijakaumaa siten, että seuraavat x-akselin pisteet sopivat niiden alle silmämääräisesti mahdollisimman hyvin:

$$(1;0),(1{,}5;0),(3;0),(4;0),(7;0),(7{,}5;0),(9;0).$$

Mitkä ovat jakaumien varianssit ja odotusarvot?

Suomen valtio on päättänyt rakentaa kolme 5G-mastoa, jotka kattavat kymmenen suurinta kaupunkia (Helsinki, Espoo, Tampere, Vantaa, Oulu, Turku, Jyväskylä, Kuopio, Lahti ja Pori). Tätä varten halutaan määrittää koordinaatit mastoille sekä saada selville, mitkä kaupungit kuuluvat mihin mastoon. Koska Suomi on niin pohjoisessa, että leveyspiirien pituus on hyvin eri kuin pituuspiirien, joudutaan leveyspiirin vaikutus skaalaamaan.

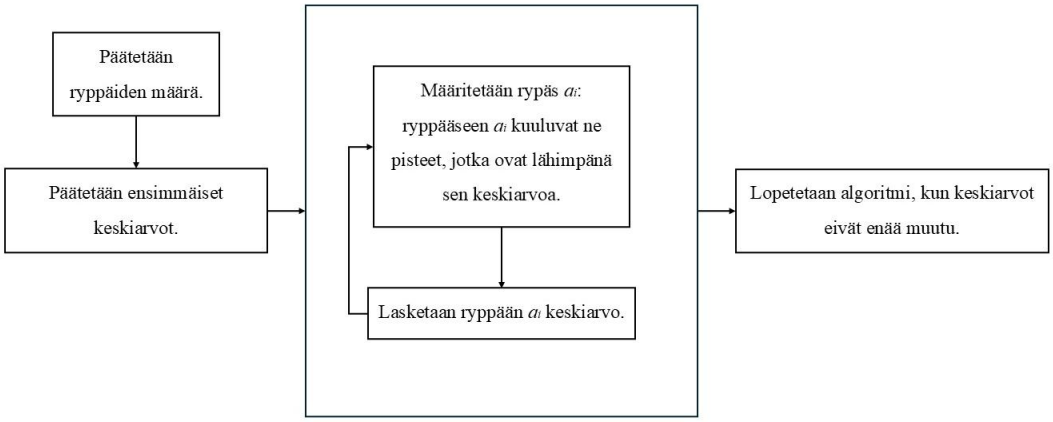
c. olettaen, että Suomi on 60° leveyspiirillä, mikä on tämän leveyspiirin pituuden ja maapallon ympärysmitan suhde?

Nimetään c-kohdasta saatu skaalausvakio kirjaimella *s*. jokainen 1 km siirtymä pohjois-etelä-akselilla vastaa siis *s* km siirtymää itä-länsi-akselilla. Koordinaattipisteiden etäisyys saadaan nyt kaavalla

$$\sqrt{(x_1 - x_2)^2 + (s \cdot (y_1 - y_2))^2},$$

jossa x-koordinaatit kuvaavat leveyspiirejä ja y-koordinaatit pituuspiirejä

d. Määritä mastojen koordinaatit alla olevalla algoritmilla lähtien listan itäisimmästä, läntisimmästä ja eteläisimmästä kaupungista. Mitkä kaupungit käyttävät samoja mastoja? (keskiarvot lasketaan laskemalla erikseen kunkin ryppään pituus- ja leveyskoordinaattien keskiarvo)



d. Mainitse kolme sovellusta, joita käytät ja jotka todennäköisesti käyttävät ohjaamatonta koneoppimista. Miten epäilet niiden keräävän dataa ja mihin tarkoitukseen?

Kesto: opetusta 30 minuuttia, tehtävän tekemistä luokassa 45 minuuttia, kotitehtävänä 2–4 päivää.

Tehtävänanto tähän tehtävään on jo huomattavasti avoimempi, koska tässä vaiheessa koneoppimisesta on tehty jo kolme tehtävää ja oppilailta voi odottaa itsenäisempää työskentelyä. Kohdissa a ja b oppilaat saavat tukea opetusosuudesta, eikä heidän tarvitse päätellä pelkistä kuvista mitä kuvissa tapahtuu. Oppilaille voi myös neuvoa normaalijakauman piirtämisen, ja selittää, että kaikkien ryppään pisteiden pitäisi olla mahdollisimman korkeassa kohdassa. On myös mahdollista, että oppilaat yrittävät tehdä tähän matemaattista tarkastelua, esim. tarkastaa, että normaalijakaumat saavat mahdollisimman suuria arvoja valituilla keskiarvoilla ja variansseilla. Tähän voi kannustaa, mutta ei voi vaatia.

C ja d kohdissa opettaja voi esitellä oppilaille kuvan 12 mukaisen hahmotelman tai Geogebra-animaation kautta, miksi leveyspiirit pitää skaalata. Kannattaa myös tarkastaa, että oppilaat saavat oikean tuloksen ($s = 0{,}5$) d-kohdassa, mutta muuten opettajan osallistuminen on hyvin riippuvainen siitä, kuinka paljon oppilailta odotetaan tiedonhakua. Opettaja voi antaa suoraan taulukon 2 mukaiset tiedot oppilaille, tai he voivat etsiä itse tiedot internetistä, jolloin opettajan kannattaa varautua selittämään kulmaminuuttien ja -sekuntien käyttö, sillä niitä käytetään laajalti. Tämä muutos minuuteista ja sekunneista kulman desimaaleihin on kuitenkin melko helppo ja taulukkolaskentaohjelma tekee laskemisesta lähes triviaalia. Oppilaille voi antaa vinkiksi käyttää karttaa apuna, jolloin oppilaiden ei tarvitse tehdä laskuja pisteille, jotka kuuluvat selvästi tiettyyn ryppääseen.

6.3.1 Ongelmalähtöinen lähestymistapa

Ongelmalähtöisessä lähestymistavassa on syytä harkita, miten oppilaille jaetaan materiaalit tehtävän kohtiin a ja b. Suoranaista opettamista ei välttämättä edes tarvita, vaan oppilaille voi jakaa itsenäisesti opeteltavat materiaalit, joita he käyttävät. Tällöin oppilaat joutuvat harjaantumaan myös tiedon lukemisessa ja tässä vaiheessa tehtäväpakettia heiltä voi vaatia myös itsenäisempää oppimista. Oppilaille on syytä antaa tehtäväksi hakea itse c-kohdan kaupungit ja koordinaatit. Tämä tuo tehtävään tiedonhakuongelman, jota tehtävässä ei muuten ole. Muuten tehtävä ei poikkea tehtävänannoltaan, sillä tehtävä on jo valmiiksi avoimempi.

6.3.2 Tehtävän purku

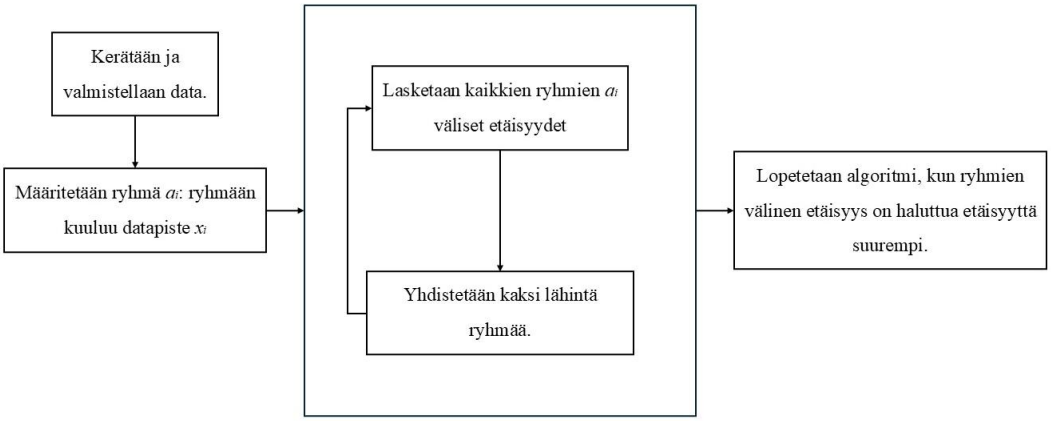
Hierarkkisen ryvästämisen vuokaavioita, josta on annettu esimerkki kuvassa 10, tarkastettaessa on syytä ottaa puheeksi menetelmän hyvät ja huonot puolet. Hierarkkinen klusterointi voidaan tehdä usealla tavalla esimerkiksi lähin naapuri tai kaukaisin naapuri- menetelmillä, mutta oikean menetelmän löytäminen voi olla hyvin vaikeaa, kun dataa ei enää pystytä piirteiden määrän takia visualisoimaan. Hierarkkinen ryvästäminen on kuitenkin varsin tehokas ja suoritettavien operaatioiden määrä pienenee jokaisella iteraatiolla. Oppilaat varmasti myös huomaavat, että algoritmi tuottaa aina vain yhden ryppään, joten todelliset ryppäät joudutaan etsimään välivaiheista. Tämä voidaan tehdä esimerkiksi katkaisemalla algoritmi tietyn etäisyyden jälkeen.

Tehtävän b-kohdassa oppilaita kannattaa myös pyytää palauttamaan kuvat heidän luomistaan käyristä, jotta vastausten tarkastaminen on helpompaa, sillä tarkkoja oikeita vastauksia ei ole, mutta oppilaiden kuvista tulisi löytyä kuvan 11 mukaiset ryppäät. Tarkka vuokaavio vaatisi hieman monimukaisempaa matematiikkaa ns. EM-algoritmi, mutta oppilaille voidaan sanoa, että tämä perustuu samanlaiseen tappiofunktion minimoimiseen kuin luokitteluongelmassa. Jakaumamenetelmän vahvuus on sen joustavuus, sillä ryppäiden välissä olevat epäselvät pisteet voivat kuulua useaan ryppääseen ja niiden todennäköisyyttä kuulua kuhunkin ryppääseen voidaan vertailla ja piste voidaan myös tarpeen

tullessa hylätä. Selvä heikkous on se, että ryppäiden määrä pitää määrätä algoritmin alussa, mikä tuottaa jälleen vaikeuksia, kun dataa ei pystytä enää visualisoimaan.

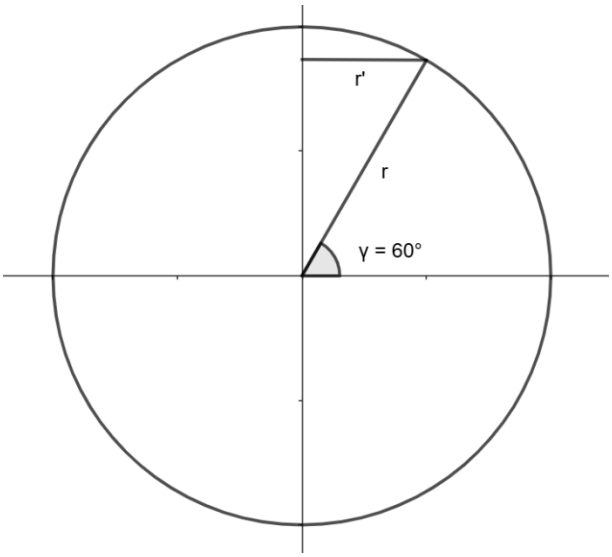
C ja d kohdissa käsitellään k-keskiarvomenetelmää, josta oppilaiden olisi hyvä huomata sen nopea suppeneminen ja yksinkertainen algoritmi, mutta myös se, että ryppäiden määrä pitää päättää algoritmin alussa. Tämän kohdan läpikäynnissä voi olla hyvä ottaa puheeksi tehtäväpaketin ensimmäinen tehtävä, jossa laskettiin keskiarvo moniulotteiselle kouludatalle, mikä näyttää, että tämä menetelmä voidaan tehdä helposti vektorimuodossa, vaikka datalla olisi huomattavan monta piirrettä. Oppilaille voi näyttää esimerkin, miten ryppäät muodostuvat eri tavalla, jos aloituspisteet valitaan toisin. Tämän takia algoritmi ajetaan useaan kertaan.

Oppilaiden kanssa kannattaa jättää aikaa pohtimaan d-kohtaa ja miten yksinkertaisia nämä menetelmät ovat ja miten paljon niitä todennäköisesti käytetään. Esimerkiksi ruokakauppaketjun etusovellukselle on helppo katsoa etukortin avulla mitä käyttäjä ostaa ja näiden ostosten perusteella profiloida käyttäjä ja syöttää tälle mainoksia. Viihteen kanssa algoritmit saattavat päättää suuren osan mitä palvelun käyttäjä näkee ja oppilailta voi kysyä esimerkiksi, kuinka moni käyttää Youtube-sovelluksessa tilausvälilehteä, jolloin he hallitsevat itse millaista sisältöä näkevät ja kuinka moni luottaa vain etusivun algoritmeihin.



Kuva 10: Esimerkki hierarkkisen ryvästyksen vuokaaviosta.

Kuva 11: Esimerkkivastaus tehtävän b-kohtaan.



Kuva 12: Havainnollistava ympyrä leveyspiirien skaalauksesta.

Kaupunki	Koordinaatit N/E	
Helsinki	60,17	24,94
Espoo	60,21	24,66
Tampere	61,50	23,76
Vantaa	60,30	25,01
Oulu	65,01	25,47
Turku	60,45	22,27
Jyväskylä	62,24	25,75
Kuopio	62,89	27,69
Lahti	60,98	25,65
Pori	61,48	21,80

Iteraatio 1

	Keskus 1 N/E		
	60,17	24,94	
	(xk-xi)^2	(yk-yi)^2	skaalattu etäisyys
Helsinki	0,00	0,00	0,00
Espoo	0,00	0,08	0,15
Tampere	1,77	1,40	1,46
Vantaa	0,02	0,00	0,14
Oulu	23,49	0,28	4,85
Turku	0,08	7,16	1,37
Jyväskylä	4,31	0,65	2,11
Kuopio	7,43	7,54	3,05
Lahti	0,66	0,51	0,89
Pori	1,73	9,91	2,05

	Keskus 2 N/E		
	62,89	27,69	
	(xk-xi)^2	(yk-yi)^2	skaalattu etäisyys
Helsinki	7,43	7,54	3,05
Espoo	7,22	9,19	3,08
Tampere	1,95	15,43	2,41
Vantaa	6,73	7,19	2,92
Oulu	4,50	4,91	2,39
Turku	5,96	29,40	3,65
Jyväskylä	0,42	3,76	1,17
Kuopio	0,00	0,00	0,00
Lahti	3,66	4,14	2,17
Pori	1,99	34,73	3,27

	Keskus 3		
	61,48	21,80	
	(xk-xi)^2	(yk-yi)^2	skaalattu etäisyys
Helsinki	1,73	9,91	2,05
Espoo	1,63	8,18	1,92
Tampere	0,00	3,86	0,98
Vantaa	1,41	10,31	2,00
Oulu	12,46	13,51	3,98
Turku	1,07	0,22	1,06
Jyväskylä	0,57	15,63	2,12
Kuopio	1,99	34,73	3,27
Lahti	0,25	14,89	1,99
Pori	0,00	0,00	0,00

Kuva 13: Esimerkki taulukko d-kohdan laskentataulukoista.

7 Tehtävä 6: suuret kielimallit

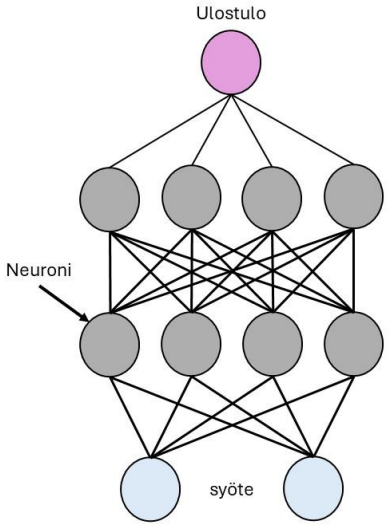
Tässä tehtävässä oppilaat tutustuvat suuriin kielimalleihin. Koska suurien kielimallien toiminta on matemaattisesti jo huomattavasti lukiomatematiikan ulkopuolella, tämän tehtävän tarkoitus on harjaannuttaa oppilaat käyttämään tekoälytyökaluja vastuullisesti ja harkitsemaan niiden käytön tarpeellisuutta sekä luotettavuutta. Oppilaille esitellään myös sanavektoreiden käsite. Tehtävässä poikkeuksellisesti ei ole eroteltu ongelmalähtöistä versiota, tai eriyttämistä, sillä tehtävänanto on luonteeltaan valmiiksi ongelmalähtöinen projekti. Tämän sijaan tehtävästä annetaan lyhyempi versio, joka on helpompi toteuttaa, mikäli aikaa on vähän.

7.1 Materiaalit

Ennen tehtävänannon jakoa oppilaille esitellään ajatus monikerroksisista koneoppimismalleista ja neuroverkoista. Tehtäväpaketin aikana on esitelty hyvin yksinkertaisia koneoppimismalleja, mutta nykyisissä koneoppimismalleissa näitä menetelmiä kerrostetaan ja yhdistellään, jotta niiden oppimiskyky saataisiin paremmaksi. Tästä voidaan ottaa esimerkiksi kuvan 14 mukainen neuroverkko, jossa useita lineaarikuvauksia ja näiden aktivointifunktioita. Jokaiseen neuroniin on liitetty oma lineaarikuvaus, joka laskee siihen tulleiden syötteiden painotetun summan, sekä epälineaarinen aktivointifunktio

Tämän lisäksi malleihin syötettävälle datalle voidaan ensin käyttää ohjaamattoman koneoppimisen keinoja, esimerkiksi luonnollista kieltä voidaan mallintaa niin sanotuiksi sanavektoreiksi, jossa sanoille voidaan määrittää moniulotteinen vektoriesitys sen käytön perusteella, esimerkiksi sanat ’’kissa’’ ja ’’koira’’ olisivat lähellä toisiaan tässä vektoriulottuvuudessa, sillä niitä käytetään usein samassa yhteydessä, kun taas ’’pingviini’’ olisi lähellä näitä eläimenä, mutta hieman kauempana, koska se ei ole kotieläin. Sanavektoreita voidaan edelleen käyttää lukemaan ja tulkitsemaan vaikka käyttäjän kysymys, sekä ennustamaan ohjatulla ja ohjaamattomalla koneoppimisella tämän kysymykseen mahdollisimman sopiva vastaus. Kun tätä vastausta hienosäädetään useiden eri koneoppimismallien ja kenties logiikkamoduulin kautta, saadaan luotua esimerkiksi hyvin ihmistä jäljittelevä tekoäly-chatbotti.

Suuret kielimallit vaativat kuitenkin kouluttamiseen valtavat määrät dataa ja niiden käyttö vaatii huomattavan paljon laskentatehoa. Vaikka ne ehkä kykenevätkin arvioimaan vastauksiaan, ne ovat silti riippuvaisia niiden koulutukseen käytettävästä datasta ja niiden parametrejä voidaan säätää suosimaan tietyn tyyppistä käyttäytymistä, esimerkiksi Open AI:n Chat-GPT ei vastaa kysymyksiin, jotka avustaisivat rikoksissa. Mikäli tekoälyllä halutaan tuottaa kuvia tai videoita, vaaditaan myös paljon valmista materiaalia, jonka käyttö koneoppimisen koulutuksessa herättää edelleen keskustelua tekijänoikeuksista.



Kuva 14: Esimerkki kaksikerroksisesta neuroverkosta.

Tehtävässä oppilaat saavat päättää itse, mistä tekoälyyn liittyvästä aiheesta tekevät esitelmän, projektin, tai vaihtoehtoisesi yksilötyönä esseen. Oppilaat saavat käyttää tehtävässä tekoälytyökaluja, mutta kaikki tekoälyn käyttö pitää dokumentoida ja kirjata ylös palautettavaan tekstitiedostoon, pois lukien esimerkiksi googlehakuun liittyvät algoritmit tai oikeinkirjoitustyökalut diaesitysohjelmassa. Taulukossa 3 on esitelty esimerkkejä projekti-, essee- tai esitelmäaiheista, mutta opettaja tai oppilaat voivat myös keksiä lisää aiheita. Tärkeää kuitenkin olisi, että aihe tukee tehtäväpaketin tavoitteita ja, että se on riittävän vaikea, että tekoälytyökalujen käyttö ei tee siitä täysin triviaalia. Koneoppimismallin koulutusprojektia ei suositella, mikäli sitä tekevillä oppilailla ei ole aiempaa harrastuneisuustaustaa ohjelmoinnista.

Taulukko 3: Esitelmä/essee aiheet.

Aihe	Projektityyppi
Tekoäly työnhaussa	Essee/esitelmä
Itseajavat autot ja tekoäly liikenteessä	Essee/esitelmä
Tekoäly valvontateknologiassa	Essee/esitelmä
Tekoälyn historia	Essee/esitelmä
Opetus ja oppiminen tekoälyn aikakaudella	Essee/esitelmä
Tekoäly ja ilmaston muutos	Essee/esitelmä
Miten tekoäly muuttaa sosiaalista mediaa?	Essee/esitelmä
Tekoälyn tuottama taide ja tekijänoikeudet	Essee/esitelmä
Tekoäly mediassa ja kulttuuri-ilmiönä	Essee/esitelmä

Tekoäly, koulutusdata ja tasa-arvo	Essee/esitelmä
Keksi ja luo tekoälyyn liittyviä matematiikan tehtäviä.	Projekti
Koodaa ja kouluta yksinkertainen koneoppimismalli.	Projekti

7.2 Tehtävänanto

Valitse/valitkaa itsellenne essee tai projektiaihe. Käyttäkää jotain tekoälytyökalua apuna ja dokumentoikaa erilliseen oppimispäiväkirjaan, esimerkiksi Word-tiedostoon, kaikki tekoälyn käyttö seuraavilla tiedoilla:

- tekoälytyökalu
- syöte
- vastaus
- Oliko vastaus oikein? tarkasta luotettavasta lähteestä!
- Mikäli vastaus oli väärin, onko tälle selkeä syy, esimerkiksi datan vähäinen määrä tai käännösvirhe?
- Voisiko syötettä muuttaa ja saada paremman vastauksen?
- Muita huomioita

Huomioi, että on huomattavasti helpompi tarkastaa pienempiä kysymyksiä ja vastauksia, kuin esimerkiksi pyytää tekoälyä kirjoittamaan pitkä teksti aiheesta ja tämän jälkeen tarkistaa kaikki faktat. Ole varovainen plagioinnin kanssa!

Kesto: 30min opetus suurista kielimalleista, 45min tehtävän jakaminen ja aloittaminen. N.1–2 viikkoa kotitehtävänä.

Riippuen ajankäytöstä ja ryhmän taidoista tehtävä voidaan antaa joko tehtäväksi ryhmissä tai yksilötöinä. Ryhminä oppilaat tekevät esitelmän, esimerkiksi 15min diaesityksen, valitsemastaan aiheesta ja samalla dokumentoivat ja arvioivat tekoälyn käyttöä projektissa. Tunnilla oppilaita voi kannustaa esimerkiksi etsimään lähteitä ja esimerkkitapauksia tekoälyn avulla. Ryhmille voi myös jakaa käytettäväksi eri tekoälytyökaluja, mutta tällöin on syytä harkita tietoturvakysymyksiä, jotka liittyvät kyseisiin palveluntarjoajiin ja vallitseviin säädöksiin.

7.2.1 Tehtävän purkaminen

Se, miten purku toteutetaan tässä tehtävässä, riippuu huomattavasti siitä, miten tehtävä on annettu. Jos tehtävä on annettu esseemuotoisena tehtävä, niin purkua ei välttämättä tarvitse toteuttaa yhteisesti, vaan oppilaan palauttama tehtävä voidaan arvioida yksilöllisesti esimerkiksi Moodlen tai vastaavan oppimisalustan kautta. Opettajan tulee harkita, miten tehtävän arvostelussa painotetaan esseen ja tekoälytyökalujen analysointia erikseen, esimerkiksi onko essee tai projekti itsessään vain formatiiviseen arviointiin, ja tekoälydokumentaatio antaa arvosanan. Täsmälliset arviointiohjeet vaatisivat tutkimusta tehtäväpaketin käytöstä, eikä tämän tutkimuksen tekeminen ole mahdollista tutkielman mittakaavassa.

Esitelmä ja projektimuodossa voidaan pitää esitelmät esim. 10–15 minuuttia, ja esitelmän lopuksi oppilaat kertovat omat kokemuksensa siitä, miten tekoälytyökalut auttoivat esitelmässä ja miten onko heidän näkökulmansa tekoälyn käyttöön muuttunut tehtävien jälkeen. Dokumentaatiot kannattaa myös palauttaa tekstimuodossa, jotta oppilaiden tekoälyn käytöstä saa laajemman kuvan kuin muutaman lauseen esitelmän lopuksi. Arviointi jää jälleen opettajan harkinnan varaan.

7.2.2 Vaihtoehtoinen tehtävä

Mikäli kurssilla ei ole aikaa tuottaa laajaa projektia, esitelmää tai esseetä, oppilaille voidaan antaa tehtäväksi lyhyt essee heidän haluamastaan aiheesta, joko aiemmalta listalta tai vaikka oppiaineiden rajoja ylittäen. Tällöin esseen pituus olisi noin puoli sivua–sivu ja oppilaat kirjoittaisivat esseen sekä täysin itse ilman tekoälyn apua, että tekoälyn avulla kotona. Esseen kirjoittamiseen saa tietenkin edelleen käyttää internetiä tietolähteenä. Oppilaita kannattaa myös kannustaa valitsemaan essee heidän omien harrastusten ja mielenkiintojensa mukaan, ja aiheiden olisi hyvä olla hieman normaalia hienostuneempia. Esimerkiksi, mikäli oppilas haluaa tehdä esseensä ratsastuksesta, esseen aihe voisi olla, vaikka ”miten hevosurheilu on muuttunut teollistumisen myötä”.

Purkutunnilla oppilaat jaetaan ryhmiin ja opettaja esittelee kullekin ryhmälle kolme esseetä, joista jokaisesta oppilaat äänestävät, onko kyseessä tekoälyn vai ihmisen tuotos sekä perustelevat vastauksensa parhaan kykynsä mukaan. Lopuksi opettajan johdolla käydään läpi ovatko oppilaat arvanneet oikein ja puretaan, miten tekoälyn tuottama teksti poikkeaa ihmisen tuottamasta. Oppilailta voidaan myös kysyä, tekivätkö tai tekisivätkö he jotain muutoksia omaan esseeseensä luettuaan tekoälyn version, esimerkiksi oliko tekoälyn tuottamassa esseessä jokin tieto tai näkökulma, joka tuli heille uutena. Näissä tilanteissa kannattaa painottaa lähdekritiikkiä, erityisesti sen suhteen, ettei kyseessä ole kielimallin hallusinaatio.

8 Loppupäätelmät

Tekoälylukutaito osana kansallista opetussuunnitelmaa on vielä vähän tutkittu, joten tarkka kirjallisuuskatsaus ja tämän tutkielman vertaaminen muiden maiden opetussuunnitelmiin on vielä tarpeen. UNESCOn teettämä kirjallisuuskatsaus ei anna riittävän tarkkaa kuvaa opetuksen sisällöstä eikä avaa tehtäviä, niiden matemaattisia tavoitteita tai niiden pedagogista pohjaa, joten se on hyödyllinen lähinnä pohtiessa tehtäväpaketin karkeita tavoitteita. Laaja kirjallisuuskatsaus useiden maiden opetusmateriaaleihin on kuitenkin tämän tutkielman ja ehkä jopa yleisesti didaktisen pro gradu-tutkielman mittakaavan ulkopuolella ennen kaikkea aineiston monikielisyyden takia.

Tehtäväpaketti vaatii myös kliinisiä tutkimuksia siitä, miten hyvin sen tavoitteet toteutuvat käytännön opetuksessa ennen, kuin se voitaisiin ottaa käyttöön. Tämä voitaisiin toteuttaa esimerkiksi tutkimalla ensin yksittäisten tehtävien tehokkuutta suljetussa ympäristössä ja tämän jälkeen tehtäväpaketin soveltuvuutta lukiokurssiympäristöön. Lukioikäisten oppilaiden ennakkotietoja ja -käsityksiä tekoälystä pitäisi myös kartoittaa, ja tutkimustietoa tarvitaan myös siitä, että onko koneoppimisen matemaattisen perustan osaaminen tehokas lisäämään tietoisuutta tekoälyn luonteesta ja yhteiskunnallisesta vaikuttavuudesta.

Tekoälytyökaluihin osana oppimista ja opetusta liittyy edelleen eettisiä kysymyksiä, mediakohua, ristiriitaista tutkimusta ja ennakkoluuloja. Tutkielman alussa lueteltujen negatiivissävytteisten lehtiartikkeleiden rinnalla on myös positiivisia tuloksia tekoälytyökalujen käytöstä osana opiskelua (Smerdon, 2024), vaikka on varmasti totta, että tekoälyä voi käyttää myös vilpillisesti. On kuitenkin selvää jo Suomeen lähivuosina sijoitettavien datakeskusten perusteella, että moderni koneoppiminen ja tekoäly ei ole katoamassa yhteiskunnasta ja myös opetusjärjestelmän täytyy olla valmis vastaamaan tähän muutokseen (Lehtilä, 2025).

9 Lähteet

Alho, J., Arjas, E., Läärä, E., & Pere, P. (2023). Tilastotieteen sanasto. *Suomen Tilastoseuran julkaisuja no. 8. 2.laitos*.

Alsayat, A., & El-Sayed, H. (2016). Social media analysis using optimized K-Means clustering. *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA).* Towson: IEEE. doi:10.1109/SERA.2016.7516129

Anderson, W., Airasian, P., Cruikshank, P., Mayer, R., Krathwohl, D., Pintrich, P., . . . Wittrock, M. (2014). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom´s.* Harlow: Pearson Education Limited.

European Union. (2024). *AI Act*. Noudettu osoitteesta European comission: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Grosz, B. J.;Grant, D. G.;Vredenburgh, K.;Behrends, J.;Hu, L.;& al., e. (2018). *Embedded EthiCS: Integrating Ethics Broadly Across Computer Science Education.* Harvard University.

Kaggle. (31. Maaliskuu 2016). Noudettu osoitteesta Titanic: https://www.kaggle.com/datasets/heptapod/titanic/data

Kaggle. (31. maaliskuu 2016). Noudettu osoitteesta European Soccer Database: https://www.kaggle.com/datasets/hugomathien/soccer

Kaggle. (2020). Haettu 31. Maaliskuu 2025 osoitteesta Ottawa Real Estate: https://www.kaggle.com/datasets/tejusrevi/ottawa-real-estate-data

Kolari, J., & Kallio, A. (2023). *Tekoäly 123: Matkaopas tulevaisuuteen.* Docendo.

Kuntaliitto. (2025). *Kaupunkien ja kuntien lukumäärät ja väestötiedot*. Haettu 8. Toukokuu 2025 osoitteesta Kuntaliitto: https://www.kuntaliitto.fi/kuntaliitto/tietotuotteet-ja-palvelut/kaupunkien-ja-kuntien-lukumaarat-ja-vaestotiedot

Kymäläinen, S. (5. helmikuuta 2025). *Helsinki kielsi opettajia käyttämästä tekoälyä – kouluissa pohditaan, saako Wordin automaattista oikolukua enää käyttää*. Noudettu osoitteesta YLE uutiset: https://yle.fi/a/74-20141409

Lehtilä, S. (14. Toukokuu 2025). *Kartta näyttää, mihin datakeskuksia nousee Suomessa*. Haettu 14. Toukokuu 2025 osoitteesta YLE: https://yle.fi/a/74-20161487

Miao, F.;& Cukurova, M. (2024). *AI competency framework for teachers.* Paris: UNESCO. doi:https://doi.org/10.54675/ZJTE2084

Opetushallitus. (2019). *Lukion opetussuunitelman perusteet 2019.* Helsinki, Suomi: Grano oy.

Poikela, S.;Vuoskoski, P.;& Kärnä, M. (2009). Developing Creative Learning Environments. Teoksessa O.-S. Tan, *Problem-based Learning And Creativity* (ss. 67-87). Singapore: Cengage Learning Asia Pte Ltd.

Slotte Dufva, T., & Mertala, P. (2021). Sähköä ja alkemiaa: Tekoälydiskurssit yleisradion verkkoartikkeleissa. *Media ja viestintä, 44(1),* 95-115.

Smerdon, D. (2024). *AI in essay-based assessment: Student adoption, usage, and performance* (Osa/vuosik. 7). Lucia, Australia: School of Economics, University of Queensland . doi:https://doi.org/10.1016/j.caeai.2024.100288.

Tan, O.-S. (2009). *Problem-based Learning And Creativity.* Singapore: Cengage Learning Asia Pte Ltd.

UNESCO. (2019). BEIJING CONSENSUS on artificial intelligence and education. *International Conference on Artificial Intelligence and Education, Planning Education in the AI Era: Lead the Leap, Beijing, 2019* (ss. 4-11). Paris: UNESCO.

UNESCO. (2022). *K-12 AI curricula A mapping of government-endorsed AI curricula.* Paris: UNESCO. doi:https://doi.org/10.54675/ELYF6010

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *UC Irwine Machine Learning Repository*. doi:10.24432/C5DW2B

Ympäristöministeriö. (2025). Haettu 9. huhtikuu 2025 osoitteesta Asuntojen hintatiedot: https://asuntojen.hintatiedot.fi/haku/

Liitteet

Liitteiden pääotsikkoa ei numeroida. Liitteissä käytetään samoja tyylejä kuin tekstiluvuissa.

Liite 1. asuntojen hinnat tehtävään 2

Sijainti	Huoneiden määrä + keittiö	Neliöt	Myyntihinta	neliöhinta	rakennusvuosi
Keskusta	1	28	110000	3929	1962
Keskusta	1	39	159000	4077	1961
Keskusta	1	33	120000	3636	1962
Keskusta	1	25	135000	5400	1961
Keskusta	1	31	134000	4323	1954
Itäinen keskusta	1	32	110000	3438	1973
Keskusta	1	26	133000	5115	1961
Keskusta	1	22	108000	4909	1956
Keskusta	1	32,5	157900	4858	2017
Itäinen keskusta	1	45	162000	3600	1931
Keskusta	1	34	137000	4029	1961
Keskusta	1	27	115000	4259	1961
Itäinen keskusta	1	27,5	118000	4291	1938
Keskusta	1	30,5	155000	5082	2020
Keskusta	1	37	158000	4270	1929
Keskusta	2	59	165000	2797	2001
Keskusta	2	59	205000	3475	1967
Keskusta	2	39	208000	5333	2016
Keskusta	2	66	245000	3712	2014
Itäinen keskusta	2	50	146500	2930	1975
Itäinen keskusta	2	38	127500	3355	1963
Keskusta	2	45	136000	3022	1957
Itäinen keskusta	2	53	163000	3075	1963
Itäinen keskusta	2	45	169000	3756	2002
Keskusta	3	97	399000	4113	1920
Keskusta	3	73,5	195000	2653	1963
Keskusta	3	77	248000	3221	1959
Keskusta	3	84	295000	3512	1990
Keskusta	3	96	272000	2833	1948
Keskusta	3	77,5	215000	2774	1912
Keskusta	3	72	244000	3389	1956
Itäinen keskusta	3	70	234000	3343	1972
Itäinen keskusta	3	68	248300	3651	1965
Itäinen keskusta	3	61	228000	3738	1955

Keskusta	3	68	202000	2971	1966
Keskusta	3	69	265000	3841	1967
Itäinen keskusta	3	70	235000	3357	1960
Keskusta	3	82	288963	3524	1962
Itäinen keskusta	3	73	210000	2877	1975
Keskusta	3	74	200000	2703	1957
Itäinen keskusta	3	83	368000	4434	1963
Keskusta	3	109	749400	6875	2021
Keskusta	3	73	348000	4767	1957
Keskusta	3	79	258500	3272	1951
Keskusta	3	71	270000	3803	1958
Itäinen keskusta	2	54	185000	3426	1958
Keskusta	3	107	270000	2523	1912
Itäinen keskusta	3	69	275000	3986	1961
Keskusta	3	79	292000	3696	1961
Keskusta	3	90	320000	3556	1995
Keskusta	3	69	230000	3333	1966
Keskusta	3	73	234000	3205	1961

Tiedot noudettu ympäristöhallituksen asuntojen hintatiedot -nettisivulta
url: https://asuntojen.hintatiedot.fi/haku/

Liite 2. Syöpäkasvainten laatu ja koko tehtävään 3

id	diagnoosi	säde (ka)	pinta-ala
842302	P	17,99	1001,00
84348301	P	11,42	386,10
843786	P	12,45	477,10
845636	P	16,02	797,80
846381	P	15,85	782,70
84799002	P	14,54	658,80
848406	P	14,68	684,50
84862001	P	16,13	798,80
8511133	P	15,34	704,40
852552	P	16,65	904,60
852631	P	17,14	912,70
852763	P	14,58	644,80
852973	P	15,30	732,40
853201	P	17,57	955,10
85382601	P	17,02	899,30
854039	P	16,13	807,20
854253	P	16,74	869,50
855133	P	14,99	698,80
855563	P	10,95	371,10

857392	P	18,22	1033,00
857793	P	14,71	656,90
859283	P	14,78	668,30
859711	H	8,89	244,00
859717	P	17,20	929,40
8610175	H	12,31	470,90
8610404	P	16,07	817,70
8610637	P	18,05	1006,00
861103	H	11,45	401,50
86135501	P	14,48	648,20
861597	H	12,36	466,10
861598	H	14,64	651,90
861648	H	14,62	662,70
861799	P	15,37	728,20
862261	H	9,79	294,50
862548	P	14,42	642,50
862722	H	6,98	143,50
862980	H	9,88	298,30
863270	H	12,36	466,70
864033	H	9,78	290,20
86408	H	12,63	480,40
864496	H	8,73	230,90
864726	H	8,95	245,20
864877	P	15,78	782,60
865128	P	17,95	982,00
865432	H	14,50	640,70
866458	H	15,10	674,50
8670	P	15,46	748,90
86730502	P	16,16	809,80
868202	P	12,77	506,30
868682	H	11,43	399,80
868871	H	11,28	384,80
869104	P	16,11	813,00
869218	H	11,43	398,00
869224	H	12,90	512,20
869254	H	10,75	355,30
869691	P	11,80	432,00
86973702	H	14,44	640,10
8711202	P	17,68	963,70
8711216	H	16,84	880,20
871122	H	12,06	448,60
87127	H	10,80	357,60
8712729	P	16,78	886,30
8712766	P	17,47	984,60
871642	H	10,66	349,60
872113	H	8,67	227,20

87281702	P	16,46	832,90
873843	H	11,41	402,00
873885	P	15,28	710,60
874662	H	11,81	428,90
874839	H	12,30	463,70
875093	H	12,77	507,40
875878	H	12,91	516,40
875938	P	13,77	588,90
877159	P	18,08	1024,00
877500	P	14,45	642,70
877989	P	17,54	951,60
87880	P	13,81	597,80
879523	P	15,12	716,60
879830	P	17,01	904,30
8810436	H	15,27	725,50
8810528	H	11,84	428,00
881094802	P	17,42	948,00
8812816	H	13,65	568,90
88143502	H	14,34	641,20
88147101	H	10,44	329,60
88147202	H	12,62	496,40
881972	P	17,05	895,00
88330202	P	17,46	920,60
88350402	H	13,64	575,30
883539	H	12,42	476,50
884437	H	10,48	337,70
884626	H	12,89	512,20
88466802	H	10,65	347,00
8860702	P	17,30	928,20
886452	P	13,96	602,40
886776	P	15,32	713,30
888264	P	17,35	933,10
888570	P	17,29	947,80
889403	P	15,61	758,60
889719	P	17,19	928,30
8910251	H	10,60	346,40
8910720	H	10,71	344,90
8910721	H	14,29	632,60
8910996	H	9,74	289,90
8911163	P	17,93	998,90
8912280	P	16,24	805,10
8912284	H	12,89	516,60
8912521	H	12,58	489,00
8913	H	12,89	515,90
8913049	H	11,26	394,10
89143602	H	14,41	651,00

891670	H	12,95	513,70
891923	H	13,77	582,70
891936	H	10,91	363,70
892214	H	14,26	633,10
89296	H	11,46	403,10
89346	H	9,00	246,30
89382601	H	14,61	664,90
89382602	H	12,76	504,10
894326	P	18,22	1027,00
894335	H	12,43	477,30
8953902	P	16,27	813,70
895633	P	16,26	826,80
896839	P	16,03	793,20
896864	H	12,98	514,00
897137	H	11,25	390,00
897374	H	12,30	464,40
89742801	P	17,06	918,60
89813	H	14,42	641,20
89827	H	11,06	373,90
898677	H	10,26	321,60
89869	H	14,76	668,70
899667	P	15,75	758,60
9010333	H	8,88	241,00
9010598	H	12,76	496,60
9010872	H	16,50	838,10
9012315	P	16,35	840,40
9012568	H	15,19	711,80
9013005	H	13,69	579,10
901303	H	16,17	788,50
9013594	H	13,66	580,60
901549	H	11,27	386,30
901836	H	11,04	372,70
90250	H	12,05	447,80
90251	H	12,39	462,90
90291	P	14,60	664,70
903011	H	11,27	392,00
90317302	H	10,26	321,60
903483	H	8,73	234,30
903507	P	15,49	744,70
904302	H	11,06	378,20
904357	H	11,80	431,90
90439701	P	17,91	994,00
904689	H	12,96	525,20
9047	H	12,94	507,60
904969	H	12,34	469,10
904971	H	10,94	370,00

905189	H	16,14	800,00
90524101	P	17,99	991,70
905501	H	12,27	466,10
905520	H	11,04	373,20
905557	H	14,99	693,70
905680	P	15,13	719,50
90602302	P	15,50	803,10
906564	H	14,69	656,10
906878	H	13,66	575,30
907145	H	9,74	289,70
907915	H	12,40	467,80
908489	P	13,98	599,50
909231	H	13,85	592,60
909410	H	14,02	606,50
909411	H	10,97	371,50
909445	P	17,27	928,80
90944601	H	13,78	585,90
9110127	P	18,03	990,00
9110732	P	17,75	981,60
9110944	H	14,80	674,80
911150	H	14,53	659,70
911201	H	14,53	644,20
911202	H	12,62	492,90
9112367	H	13,21	537,90
9112712	H	9,76	290,90
911296201	P	17,08	930,90
9113156	H	14,40	646,10
9113538	P	17,60	980,50
9113846	H	12,27	465,40
911916	P	16,25	815,80
91227	H	13,90	602,90
912600	H	15,73	747,20
913102	H	14,64	666,00
913535	P	16,69	857,60
91376701	H	12,25	466,50
91376702	H	17,85	992,10
914062	P	18,01	1007,00
914366	H	12,65	485,60
915452	H	16,30	819,80
915460	P	15,46	731,30
915664	H	14,81	680,70
915940	H	14,58	658,80
917080	H	12,75	493,80
91789	H	11,26	388,10
917897	H	9,85	293,20
91813702	H	12,34	468,50

918192	H	13,94	594,20
91903902	H	13,68	575,50
919537	H	10,96	365,60
91979701	P	14,27	629,80
921092	H	7,73	178,80
921362	H	7,69	170,40
921386	H	14,47	656,40
921644	H	14,74	668,60
922296	H	13,21	538,40
922577	H	10,32	324,90
922840	H	10,26	320,80
924084	H	12,77	507,90
924632	H	12,88	514,30
924934	H	10,29	321,40
925277	H	14,59	657,10
925622	P	15,22	716,90
926954	P	16,60	858,10