# Tiedonlouhinta terveydenhuollossa, sosiaalisessa mediassa ja koulutuksessa: Mahdollisuudet ja haasteet

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tieto- ja viestintätekniikka
Toukokuu 2024
Elias Viitanen

TURUN YLIOPISTO
Tietotekniikan laitos

Elias Viitanen: Tiedonlouhinta terveydenhuollossa, sosiaalisessa mediassa ja koulutuksessa: Mahdollisuudet ja haasteet

TkK-tutkielma, 23 s.
Tieto- ja viestintätekniikka
Toukokuu 2024

---

Tässä kandidaatintyössä tutkitaan tiedonlouhinnan sovelluksia eri aloilla, keskittyen erityisesti terveydenhuoltoon, sosiaaliseen mediaan ja koulutukseen. Tutkimus selvittää, miten tiedonlouhinta voi tarjota arvokkaita oivalluksia ja edistää yhteiskunnallista kehitystä ja hyvinvointia näillä aloilla.
Taustaluvussa esitellään tiedonlouhinnan perusteet ja yleisimmät menetelmät, luoden vankan perustan sen toiminnan ymmärtämiselle ja merkitykselle eri konteksteissa.
Seuraavissa luvuissa tutkitaan tiedonlouhinnan sovelluksia terveydenhuollossa, sosiaalisessa mediassa ja koulutuksessa, analysoiden kunkin alan haasteita ja mahdollisuuksia. Terveydenhuollossa tiedonlouhinnalla on lupausta parantaa potilaiden hoitoa, resurssien kohdentamista ja tautien ennaltaehkäisyä. Sosiaalisessa mediassa se mahdollistaa suurten datamäärien analysoinnin ihmisten kommunikaation, vuorovaikutusmallien ja alustojen kehittämisen ymmärtämiseksi. Koulutuksessa tiedonlouhinta helpottaa yksilöllistä oppimistuen tarjoamista, politiikan muotoilua ja eettisiä näkökohtia koskien tietosuojaa ja datan käyttöä.
Tutkimus korostaa eettisen ja vastuullisen datankäsittelyn merkitystä, erityisesti koulutusdatan louhinnan yhteydessä.
Yhteenvetona voidaan todeta, että jatkuva tiedonlouhinnan ja oppimisanalytiikan kehitys ja soveltaminen tarjoavat lupaavia mahdollisuuksia ihmiskäyttäytymisen ja tarpeiden parempaan ymmärtämiseen eri aloilla, päätöksentekoprosessien parantamiseen ja yleiseen tehokkuuteen.

Asiasanat: Tiedonlouhinta, Terveydenhuolto, Sosiaalinen media, Koulutus, Oppimisanalyytikka, Tietotekniikka

# Sisällys

# Kuvat

# Taulukot

# 1 Johdanto

Verkossa tuotetaan suuria määriä tietoa päivittäin. Tietoa on monipuolisesti ja se jakautuu eri aloille, kuten terveydenhuolto, sosiaalinen media ja koulutus. Tämä tiedon määrä tarjoaa valtavasti mahdollisuuksia ymmärtää paremmin ihmisten käyttäytymistä, ennustaa tulevia trendejä, parantaa päätöksen tekoa ja muuten tehostaa prosesseja. Tiedonlouhinta (engl. Data mining) on menetelmä, joka mahdollistaa arvokkaiden tietojen erottamisen suurista tietomassoista luettavampaan muotoon.

Tiedonlouhinnalla on monia sovelluksia eri aloilla. Esimerkiksi terveydenhuollossa se voi auttaa tunnistamaan potilaiden riskitekijöitä, parantamaan diagnooseja ja ennustamaan sairauksien leviämistä. Sosiaalisessa mediassa se voi auttaa yrityksiä ymmärtämään paremmin asiakkaidensa mieltymyksiä ja käyttäytymistä sekä tunnistamaan trendejä ja vaikuttajia. Koulutuksessa se voi tukea opettajia ja oppilaita oppimisprosessissa, tarjota räätälöityä oppimateriaalia ja arvioida opetuksen tehokkuutta.

Tiedonlouhinnan merkitys korostuu erityisesti digitalisaation aikakaudella, jossa tieto muodostaa olennaisen osan liiketoimintaa, tutkimusta ja päätöksentekoa. Tehokkaalla tiedonlouhinnalla voidaan tuottaa arvokasta tietoa, joka auttaa organisaatioita ja yhteiskuntaa tekemään informoituja päätöksiä, parantamaan prosessejaan ja innovoimaan uusia ratkaisuja. Se tarjoaa mahdollisuuden syventää ymmärrystämme maailmasta ja hyödyntää tätä ymmärrystä tehokkaammin ja kestävämmin.

Tässä kandidaatintutkielmassa keskitymme tutkimaan ja analysoimaan tiedon-

louhinnan menetelmiä ja sovelluksia eri aloilla, erityisesti terveydenhuollossa, sosiaalisessa mediassa ja koulutuksessa. Näillä aloilla tiedonlouhinnalla on valtava potentiaali tarjota arvokasta tietoa, joka voi edistää yhteiskunnan kehitystä ja hyvinvointa. Tutkimuskysymykset ovat:

1. Mitkä ovat yleisimmät tiedonlouhinnan menetelmät ja tekniikat?
2. Miten tiedonlouhinta on sovellettavissa terveydenhuollon, sosiaalisen median ja koulutuksen alueilla?
3. Millaisia haasteita ja mahdollisuuksia tiedonlouhintaan liittyy eri aloilla?

Tutkielman tiedonhankintaan käytettiin systemaattista lähestymistapaa, joka alkoi tarkkaan haku- ja rajausstrategioiden määrittelyllä. Alkuvaiheessa pääasialliset hakukoneet ja tietokannat, kuten Volter, ProQuest, Web of Science ja Google Scholar, valittiin tiedonlouhintaan liittyvien aiempien tutkimusten tunnistamiseksi. Hakuprosessi alkoi laajalla hakusanajoukolla, joka sisälsi termejä kuten "Data mining"ja "Artificial Intelligence", ja ajanjakson rajaamisella poissulkeakseen ennen vuotta 2010 julkaistut teokset.

Koska alkuperäinen hakuprosessi tuotti runsaasti tuloksia, hakusanoja hienosäädeltiin ja rajattiin yhä tarkemmin vastaamaan tutkielman erityistä aihetta ja alueita, kuten terveydenhuoltoa, sosiaalista mediaa ja koulutusta. Tämä tarkempi lähestymistapa auttoi keskittymään olennaiseen tietoon ja minimoimaan ei-olennaisten tulosten määrän.

Tutkimusosassa käytetyt hakusanat olivat ("Data mining"OR Knowledge discovery) AND (AI OR "Artificial Intelligence") AND (Social media/Healthcare/Educational) ja julkaisuvuosien rajaukset muuttuivat hakujen mukaan. Tämä lähestymistapa varmisti, että löydetty tieto oli ajankohtaista ja relevanteinta tutkielman kannalta.

Tiedonhaun jälkeen tulokset analysoitiin ja arvioitiin niiden relevanssin ja luotettavuuden perusteella. Lopullisessa tutkielmassa käytetty aineisto oli valittu huolellisen arvioinnin jälkeen, mikä tuki tutkielman luotettavuutta.

Tutkielman rakenne on suunniteltu tarjoamaan systemaattinen lähestymistapa tiedonlouhinnan menetelmien ja sovellusten tutkimiseen eri aloilla. Kukin osa tukee tutkielman tavoitteita ja vastaa keskeisiin tutkimuskysymyksiin.

Toisessa luvussa esitellään tiedonlouhinnan perusteet ja yleisimmät menetelmät. Tässä osassa tarkastellaan erilaisia tiedonlouhinnan tekniikoita, kuten klusterointi, luokittelu ja assosiaatiosäännöt, sekä niiden perusperiaatteita ja sovellusalueita. Tämän pyrkimys on luoda vankka perusta ymmärtää tiedonlouhinnan toimintaa ja sen merkitystä eri konteksteissa.

Kolmannessa luvussa keskitytään tiedonlouhinnan sovelluksiin terveydenhuollon, sosiaalisen median ja koulutuksen aloilla. Jokaisessa osassa analysoidaan, miten tiedonlouhintamenetelmiä voidaan soveltaa kyseisellä alueella ja mitä hyötyjä ja haasteita niiden käytössä voi olla. Esimerkkeinä voidaan mainita terveydenhuollon potilastietojen analysointi, sosiaalisen median käyttäjädatan segmentointi ja koulutuksellisten trendien ennustaminen.

Yhteenvedossa tarkastellaan saatuja tuloksia ja niiden merkitystä tutkielman tavoitteiden kannalta. Analyysin kautta pyritään löytämään yhteyksiä tiedonlouhinnan menetelmien ja sovellusten välillä eri aloilla sekä arvioimaan niiden vaikutusta yhteiskunnan kehitykseen ja hyvinvointiin. Lisäksi pohditaan mahdollisia jatkotutkimuksen suuntia ja kehitysmahdollisuuksia tiedonlouhinnan alalla.

# 2 Tausta

## 2.1 Tiedonlouhinta

Tietokantojen runsas lisääntyminen lähes kaikilla tieteenaloilla loi huomattavan tarpeen tehokkaille välineille, jotka mahdollistaisivat suuren raakadatamäärän muuttamisen hyödylliseksi ja tavoitehakuisesti suunnatuksi tiedoksi. Tämän kasvaneen tarpeen tyydyttämiseksi tutkijat alkoivat tutkia koneoppimisen kehittämiä ideoita ja menetelmiä. Tällaisia menetelmiä ovat muun muassa kuviotunnistus(engl.pattern recognition), tilastollinen data-analyysi (engl.statistical data analysis), datan visualisointi (engl.data visualization), neuroverkot (engl.neural networks) ja muut vastaavat. Näiden ponnistelujen tuloksena syntyi uusi tutkimusalue, joka tunnetaan yleisesti tiedonlouhintana ja tiedon löytämisenä (knowledge discovery in data). [1]

Tiedonlouhinta on vakiintunut termi tietojenkäsittelytieteen alalla ja sen alkuperät voidaan jäljittää 1980-luvulle, jolloin termi otettiin käyttöön tutkimusyhteisössä. Alkuvaiheissa tiedonlouhinnan käsitteestä vallitsi yhteisymmärrys siitä mitä se tarkalleen merkitsi ja osittain tämä yhteisymmärrys säilyy edelleen. Yleisesti ottaen tiedonlouhinta voidaan määritellä joukoksi mekanismeja ja tekniikoita, jotka on toteutettu ohjelmistoilla ja joiden tarkoituksena on paljastaa piilotettua informaatiota datasta. Tässä yhteydessä "piilotettu"on keskeinen käsite. Vaikka SQL-kyselyt voivat olla edistyneitä, ne eivät itsessään ole tiedonlouhintaa. On myös tärkeää tulkita termi "informaatio" laajassa merkityksessä. 1990-luvun alkuun mennessä tiedonlou-

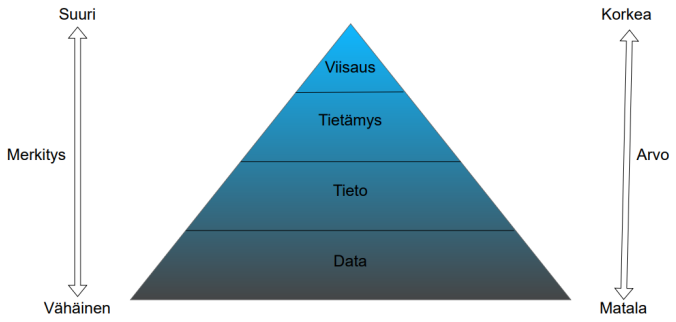
hinta tunnistettiin yleisesti aliprosessina laajemmassa prosessissa, jota kutsuttiin tiedonhauksi tietokannoista eli "KDD"(engl. Knowledge Discovery in Databases). Nykyaikaisessa tiedonlouhinnassa on kuitenkin perustellumpaa käyttää termiä "Tiedon löytäminen datasta", koska se huomioi sen, että emme enää rajoitu pelkästään tietokantoihin. [2]

Tiedonlouhintatutkimuksen tilastollisesta näkökulmasta voidaan havaita, että erityispiirteenä on laskennallisten analyysien merkityksen korostaminen perinteisten tilastollisten käsitteiden yhdistämisessä. Näihin tilastollisiin käsitteisiin kuuluvat riittävät tilastolliset menetelmät, todennäköisyydet ja mallidiagnostiikka. Tämä laskennallinen lähestymistapa, jota voidaan kutsua "laskennalliseksi kulttuuriksi"(engl. computational culture), on suoraa seurausta siitä, että tiedonlouhinta on ollut pääosin tietojenkäsittelytieteilijöiden hallitsemaa. Tietojenkäsittelytieteen kaksi merkittävää ala-aluetta, koneoppiminen ja tietokannat, ovat vaikuttaneet eniten tiedonlouhintatutkimuksen kehittymiseen viimeisten 10 vuoden aikana. [3]

Perinteiset tilastotieteelliset menetelmät ovat historiallisesti tarjonneet malleja, jotka perustuvat enemmän tai vähemmän yksityiskohtaisiin oletuksiin datan jakautumisesta. Bayesiläisen päättelyn klassinen teoria on osoittanut hyödyllisyytensä monilla sovellusalueilla kattamalla laaja-alaisesti niin lääketieteelliset sovellukset kuin kuluttajatietojen analyysin ja markkinakoron tutkimuksen. Lisäksi hermoverkot ja koneoppiminen ovat tuoneet uusia näkökulmia, käsitteitä ja algoritmeja näiden tietojoukkojen analysointiin innovatiivisella tavalla. Bayesilainen päättely on tilastollinen menetelmä, joka mahdollistaa todennäköisyyksien päivittämisen uuden tiedon saapuessa ja se on erityisen hyödyllinen monimutkaisten riippuvuuksien analysoinnissa suurista tietojoukoista. Uudet lähestymistavat, jotka ovat nousseet esiin viime vuosikymmeninä, eroavat perinteisestä tilastollisesta data-analyysistä monin tavoin. Nämä lähestymistavat luottavat vähemmän oletuksiin tiedon todellisesta jakaantumisesta ja välttävät yksinkertaisia analyyttisiä malleja. Tämän sijaan ne hyö-

dyntävät monimutkaisia malleja, jotka kykenevät oppimaan monimutkaisia epälineaarisia riippuvuuksia suurista tietojoukoista. Perinteiset tilastomenetelmät toimivat usein teoriapohjaisina työkaluina hypoteesien testaamiseen. Toisin kuin ne, koneoppiminen ja hermoverkkotutkimus arvioivat suorituskykyään sen perusteella, kuinka hyvin nämä pystyvät yleistämään uuteen dataan. Tämä data on peräisin samasta tuntemattomasta prosessista, joka on tuottanut koulutusdatan. Yleistyssuorituskyvyn arviointi eroaa merkittävästi laajalle levinneestä, mutta kyseenalaisesta käytännöstä, jossa 'kidutetaan dataa, kunnes se tunnustaa'. [1]

Kuvassa 2.2 on viisauden hierarkia-pyramidi. Data-tieto-tietämys-viisaus hierarkiaa, jota kutsutaan eri nimillä "tietohierarkia", "tietopyramidi", "viisauden hierarkia-pyramidi" on yksi tietokirjallisuuden perusmalleista, laajalti tunnustetuista ja "itsenäisiksi katsotuista" malleista. Sitä lainataan tai käytetään epäsuorasti usein data-tiedon määritelmissä tiedonhallinnan ja tietojärjestelmien kirjallisuudessa. Tyypillisesti tieto määritellään datana, tietämys tietona ja viisaus tietämyksenä. Vähemmän yksimielisyyttä on kuvattaessa prosesseja, jotka muuttavat hierarkiassa alempana olevia elementtejä niiden yläpuolella oleviksi, mikä johtaa määritelmän selkeyden puutteeseen.[4] Data on symboleita, jotka edustavat objektien ja tapahtumien ominaisuuksia. Informaatio koostuu käsitellystä datasta, jossa pyritään lisäämään sen hyödyllisyyttä. Esimerkiksi dataa kerätään kyselyiden tai mittauslaitteiden avulla ja vaestönlaskijat keräävät dataa, jonka jälkeen tilastoviranomainen muuntaa sen informaatioksi, joka esitellään taulukoina. Tietämys liittyy ohjeisiin ja vastauksiin "kuinka tehdä" -kysymyksiin. Tietämys mahdollistaa tehokkuuden lisäämisen, mutta se ei vielä käsittele syvällisempää ymmärrystä. Viisaus liittyy arvoihin ja harkintaan. Se liittyy arviointiin siitä, mikä on tehokasta tietylle tavoitteelle tai arvokkaalle päämäärälle. [5]




Kuva 2.1: Data-Viisaus diagrammi

[4]

Tietojenlouhinnasta saatujen käyttökelpoisen datan ja pätevien kaavojen löytämiseksi on monimutkaista ja tuottaa lukuisia ongelmia. Nämä haasteet kietoutuvat tietoihin, työkaluihin ja tekniikoihin, turvallisuuskysymyksiin, tulosten esittämiseen ja visualisointiin, tiedon fuusiointiongelmiin sekä institutionaaliseen sitoutumiseen ja rahoitukseen.[6]

Taulukko 2.1 esittelee tiedonlouhinnan haasteita, jotka ovat olennainen osa tiedonlouhinnan prosessia. Nämä haasteet vaihtelevat tietojen laadusta ja integroinnista suurten tietojoukkojen käsittelyyn ja monimutkaisten tietorakenteiden hallintaan. Tiedonlouhintaan liittyvien haasteiden ymmärtäminen on keskeistä tehokkaan ja luotettavan tiedonlouhinnan varmistamiseksi.

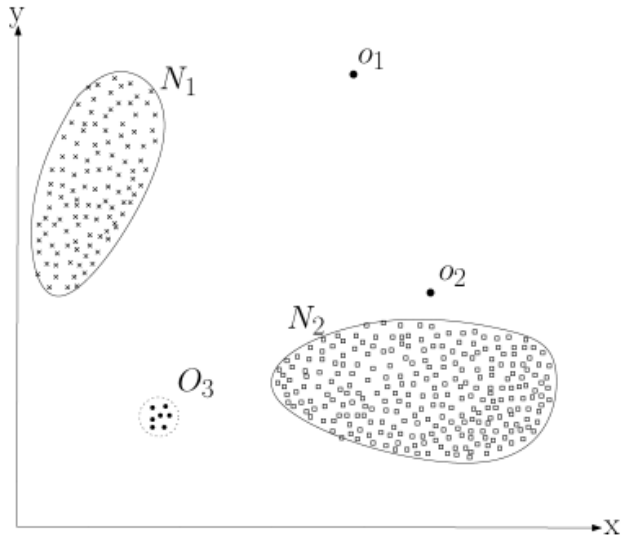
Taulukko 2.1: Tiedonlouhinnan haasteet [6]

| **Haaste** | **Kuvaus** |
|---|---|
| 1. Tietojen laatu | Epätarkka tieto, puuttuvat tai virheelliset arvot, riittämätön aineistokoko, huono tietojenkeruu. Vaatii puhdistus- ja analysointimenetelmiä. |
| 2. Tietojen integrointi | Ristiriitaiset tai päällekkäiset tiedot eri lähteistä ja muodoista: multimedia, paikkatiedot, teksti, sosiaaliset ja numeeriset tiedot. |
| 3. Tietojen saatavuus | Vaikea pääsy tai puute tietoihin. |
| 4. Suurten tietojoukkojen käsittely | Vaatii hajautettuja lähestymistapoja. |
| 5. Ei-staattiset, epätasapainoiset ja kustannusherkät tiedot | Vaatii erityiskäsittelyä. |
| 6. Tietojen kaivaminen heterogeenisistä lähteistä | Tietokantojen hakeminen eri tietolähteistä LAN- ja WAN-verkoista. Rakenteet voivat olla järjestettyjä tai puolijärjestettyjä |
| 7. Monimutkaisten ja rakenteettomien tietojen käsittely | Vaatii muuntamista rakenteelliseen muotoon. |

## 2.2 Poikkeamien havaitseminen

Poikkeamien havaitseminen on käytössä merkittävien muutosten löytämiseksi tietojoukoista [7]. Poikkeamilla tarkoitetaan datan kuvioita, jotka eivät noudata hyvin määriteltyä normaalikäyttäytymisen käsitettä. Kuvassa 2.2 on havainnollistettu poikkeamia yksinkertaisessa kahden ulottuvuuden tietoaineistossa. Datalle on kaksi normaalia aluetta, N1 ja N2, sillä suurin osa havainnoista sijoittuu näihin alueisiin.. Pisteet, jotka ovat riittävän kaukana näistä alueista, kuten pisteet o1 ja o2 sekä pisteet alueella O3 ovat poikkeamia. [8]

Tietojen poikkeamia voi esiintyä monista syistä. Näitä syitä voivat olla esimerkiksi ilkeämielinen toiminta, kuten luottokorttipetokset, kyberhyökkäykset, terroristitoiminta tai järjestelmän vikaantuminen. Kaikilla näillä syillä on kuitenkin yksi yhteinen piirre: ne ovat analyytikoille kiinnostavia. Poikkeamien kiinnostavuus tai todellinen elämän relevanssi on poikkeamien havaitsemisen keskeinen piirre. [8]



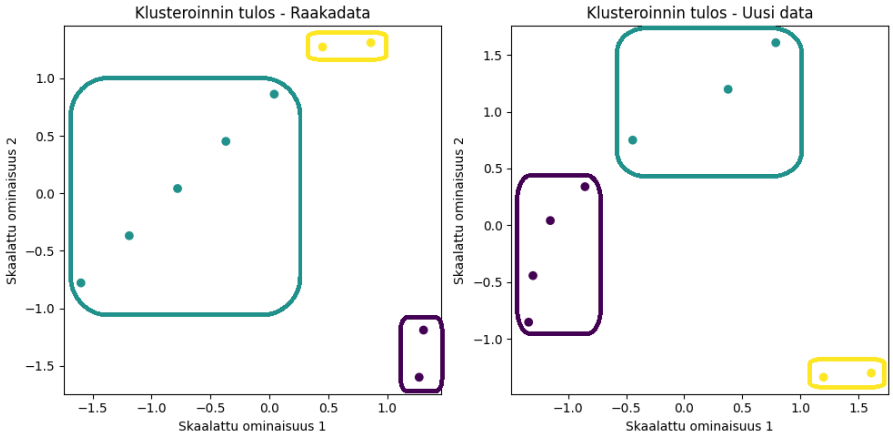

Kuva 2.2: Yksinkertainen esimerkki poikkeamiesta kaksiulotteisesta tietoaineistosta [8]

## 2.3 Klusterointi

Klusterointi on koneoppimisen menetelmä, joka ryhmittelee samankaltaisia aineistoja yhteen. Tavoitteena on luokitella data niin, että samassa ryhmässä olevat kohteet ovat keskenään samankaltaisempia kuin eri ryhmissä olevat kohteet. Tämä auttaa löytämään rakenteita ja piirteitä datasta, jotka voivat olla vaikeasti havaittavissa muuten. Klusterointia käytetään usein tietojen segmentoinnissa ja ryhmittelyssä, mikä auttaa analysoimaan suuria tietomääriä tehokkaasti.[9]

Suoritin klusteroinnin kahdelle tietojoukolle, nimittäin raakadatalle ja uudelle datalle, jotta voisimme havainnollistaa klusteroinnin toimintaa. Kuvassa 2.3 ensimmäisen klusteroinnin tuloksena esitetään, miten data ryhmitellään erillisiksi klustereiksi. Nämä havainnot ja ominaisuudet ovat haasteellisia havaita ennen klusterointianalyysiä. Klusteroinnin toteutuksessa hyödynnetään ominaisuuden normalisointia, joka perustuu ominaisuuksien keskiarvon normalisointiin ja skaalaamiseen yksikkövaihteluksi. Tällä lähestymistavalla ohjelma oppii ensimmäisestä klusteroinnista, mikä johtaa tarkempaan toiseen klusterointiin. Uuden datan klusterijakaumasta

voidaan päätellä klusterointimallin yleistettävyyttä.




Kuva 2.3: Klusteroinnintulos

Klusteroinnissa tietoja ryhmitellään luokkiin. Tämä on erityisen toivottavaa asiakastiedoissa, joissa on hyödyllistä ryhmitellä samankaltaisia asiakkaita esimerkiksi kohdennettua mainontaa varten. Tyypillisesti haluamme klusteroida tiedot joko tiettyyn määrään klustereita, kuten paremmin tunnetun K-means-algoritmin tapauksessa.[2] Vaihtoehtoinen lähestymistapa on ottaa käyttöön hierarkkinen klusterointi, jossa dataa jaetaan iteratiivisesti muodostaen joukon klustereita. Hierarkkisen klusteroinnin yleisimmin käytettyä algoritmia voidaan todennäköisesti pitää BIRCH-menetelmänä. [10] Klusterikonfiguraation "hyvyys" mitataan yleensä klusterin sisäisen yhtenäisyyden ja klustereiden välisten etäisyyksien perusteella [2]. Klusterien välinen etäisyys kuvaa klusterointianalyysissä klusterien välisiä etäisyyksiä. Se mittaa klusterien välisen erottuvuuden astetta datassa. Suurempi etäisyys indikoi selkeämpiä eroja klustereiden välillä, mikä voi viitata selkeämpiin ryhmiin tai rakenteisiin datassa. Toisaalta pienempi etäisyys voi viitata siihen, että klusterit ovat lähempänä toisiaan tai jopa sekoittuvat keskenään, mikä voi osoittaa datan monimutkaisempaa rakennetta tai epäselvempiä ryhmiä.[11] Ongelma, joka liittyy vakiintuneisiin klusterointi-algoritmeihin, kuten K-means ja KNN (K-nearest neighbors), on se, että luodut klusterit esitetään usein hyperpalloina, vaikka tämä ei aina ole

ihanteellinen muoto [2]. Hyperpallon käsite viittaa geometriseen muotoon moniulotteisessa avaruudessa, jossa kaikki pisteet ovat tietyn etäisyyden päässä keskipisteestä. Tällaiset klusterit eivät välttämättä vastaa datajoukon todellista rakennetta, mikä voi johtaa epätarkkaisiin tai epätyydyttäviin klusterointituloksiin.[12] Lisäksi haasteita liittyy usein syötedatan suureen ulottuvuuteen ja kohinaan (outliereihin) sekä kategoriseen dataan [2].

Klusterointi on vakiintunut tiedonlouhinnan (ja sitä ennen koneoppimisen) tekniikka. Mielenkiintoista on, ettei ole olemassa "parasta" klusterointialgoritmia, joka sopisi kaikelle datalle, sen sijaan jostain syystä jotkut algoritmit toimivat paremmin tietyillä datamäärillä kuin toiset. [2]

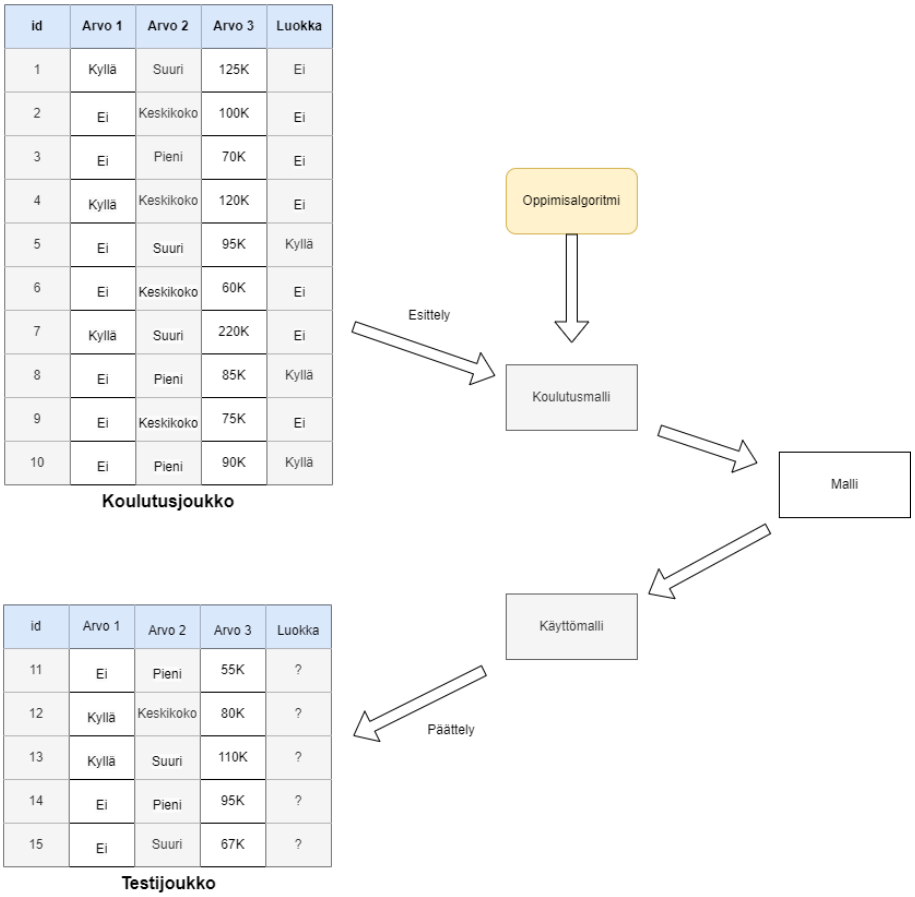
## 2.4 Luokittelu

Luokittelu (engl. Classification) on koneoppimisen osa-alue, jossa pyritään ennustamaan esimerkkitietojen avulla kohteiden luokka tai ryhmä. Tavoitteena on oppia malli, joka pystyy luokittelemaan uusia esimerkkejä niiden ominaisuuksien perusteella oikeaan luokkaan. Luokittelua sovelletaan yleisesti valvotussa oppimisessa, jossa koulutusdatassa jokaiselle esimerkille on määritelty siihen liittyvä luokka. Mallin tavoitteena on oppia tunnistamaan piirteitä, jotka erottavat eri luokat toisistaan, ja yleistää tämä tieto uusien, tuntemattomien esimerkkien luokitteluun. [11]

Luokittelua käytetään jokaisen tiedon kohteen luokitteluun tietojoukossa yhteen ennalta määriteltyyn luokkaan tai ryhmään. Tietojen analyysitehtävässä luokittelu on tilanne, jossa malli tai luokittelija rakennetaan ennustamaan kategorisia tunnisteita(luokkatunnisteiden arvoja). Luokittelu on tietojenlouhinnan toiminto, joka määrittelee kohteet kokoelmissa kohdekategorioihin tai luokkiin. Luokittelun tavoitteena on ennustaa kohde luokka tarkasti jokaiselle tapaukselle tiedoissa. Esimerkiksi luokittelumallia voitaisiin käyttää tunnistamaan lainanhakijat alhaisiksi, keskitasoisiksi tai suuriksi luottoriskeiksi. Luokittelu tehtävä alkaa tietojoukosta, jossa luok-

kien määritykset ovat tiedossa. Esimerkiksi luottoriskin ennustava luokittelumalli voitaisiin kehittää perustuen havaittuihin tietoihin monista lainanhakijoista tietyn ajanjakson aikana. Lisäksi omistus- tai vuokraustilanteet, asumisvuodet, sijoitusten määrän ja tyypin jne. Luottoluokitus olisi kohde, muut attribuutit olisivat ennustajia ja jokainen asiakas muodostaisi tapauksen.

Luokittelut ovat diskreettejä eivätkä sisällä järjestystä. Jatkuvat liukulukuarvot osoittavat numeerisen, eivätkä kategorisen kohteen. Ennustemalli, jolla on numeerinene kohde, käyttää regressioalgoritmia, ei luokittelu-algoritmia. Yksinkertaisin luokitteluongelman tyyppi on binääriluokittelu. Binäärisessä luokittelussa kohdeattribuutilla on vain kaksi mahdollista arvoa: esimerkiksi korkea luottoluokitus tai matala luottoluokitus. Moni luokkaisissa kohteissa on yli kaksi arvoa: esimerkiksi matala, keskitaso, korkea tai tuntemation luottoluokitus. Mallin rakentamisessa (koulutuksessa) luokittelu algoritmi etsii suhteita ennustajien arvojen ja kohteen arvojen välillä. Eri luokittelualgoritmit käyttävät erilaisia tekniikoita suhteide löytämiseen. Nämä suhteet tiivistetään malliin, jota voidaan sitten soveltaa erilaiseenn tietojoukkoon, jossa luokkien määritykset ovat tuntemattiomia. Luokittelulla on monia sovelluksia asiakassegmentoinnissa, liiketoimintamallinnuksessa, markkinoinnissa, luottotarkasteluissa, lääketieteellisessä ja lääkevasteiden mallinnuksessa. Tietojen luokittelu määritellään kaksivaiheisena prosessina kuvan 2.4 osoittamalla tavalla. [13]




Kuva 2.4: Luokittelu prosessin mallintaminen

[13]

## 2.5 Yhdistäminen

Yhdistäminen (engl. Association) on tietojen analyysin osa-alue, joka pyrkii tunnistamaan mielenkiintoisia suhteita ja yhteyksiä tietojoukoissa. Tämä menetelmä tutkii, miten erilaiset muuttujat esiintyvät yhdessä tai toistensa kanssa ja pyrkii löytämään säännönmukaisuuksia näiden esiintymisten välillä. Yhdistämisen avulla voidaan paljastaa esimerkiksi tuotteiden välisiä suhteita kauppa-analytiikassa, asiakkaiden ostokäyttäytymistä, tai vaikkapa säännönmukaisuuksia sairauksien ja niiden riskitekijöiden välillä terveystutkimuksessa. [14]

Tärkeimmät yhdistämisen menetelmät ovat Apriori-algoritmi ja FP Growth -algoritmi, jotka perustuvat erilaisiin lähestymistapoihin yhdistämisen ongelman ratkaisemisessa. Apriori-algoritmi käyttää apriori-periaatetta, jonka mukaan jos jokin yhdistelmä on yleinen, niin sen osajoukot ovat yleisiä. FP Growth -algoritmi puolestaan hyödyntää puurakennetta ja tiivistettyä tallennusta yhdistämisen tehokkaampaan toteuttamiseen. [14]

# 3 Tiedonlouhinta eri aloilla

## 3.1 Tiedonlouhinta terveydenhuollossa

Terveydenhuolto kattaa yksityiskohtaiset prosessit sairauksien, vammojen ja muiden fyysisten tai henkisten vaivojen diagnosoinnissa, hoidossa ja ennaltaehkäisyssä ihmisillä [15]. Terveydenhuoltoala kehittyy useimmissa maissa nopeasti. Terveydenhuoltoala voidaan nähdä alana, jossa on runsaasti tietoa, koska siellä syntyy valtavia määriä tietoa, mukaan lukien sähköiset potilastiedot, hallinnolliset raportit ja muut vertailulöydökset [16]. Nämä terveyshuollon tiedot ovat silti vajaakäytössä. Tiedonlouhinnalla voidaan etsiä uutta ja arvokasta tietoa näistä suurista tietomääristä. Tiedonlouhintaa terveydenhuollossa käytetään pääasiassa erilaisten sairauksien ennustamiseen sekä avustamaan lääkäreitä kliinisten päätöstensä tekemisessä diagnoosien osalta. Tiedonlouhinta menetelmiä, joita terveydenhuoltoalalla käytetään ovat poikkeamien havaitseminen, klusterointi ja luokittelu. [17]

Lääketieteelliset tiedot generoidaan pääasiassa potilaiden hoidon kautta. Siksi lääketieteellisen tiedonlouhintaan liittyy väistämättä yksityisyys- ja oikeudellisuus kysymyksiä. Tästä syystä tietojenlouhinta biolääketieteen ja terveydenhuollon aloilla eroaa merkittävästi muilla aloilla tehdystä tiedonlouhinnasta. Tämä keskeinen ero vaatii keskustelua biolääketieteen ja terveydenhuollon alojen tietojenlouhinnan ainutlaatuisuudesta. [18]

Terveydenhuollon alalla tietojenlaatu on heikompi kuin muilla aloilla monista syistä johtuen:

1. Lääketieteelliset tiedot sisältävät väistämättä paljon puuttuvia arvoja [19]. Tämä johtuu siitä, että jopa saman sairauden omaavilla potilailla ei aina ole identtisiä tutkimuksia ja laboratorio kokeita (eri ikä, oireet, perhehistoria tai komplikaatioriskit voivat vaihdella), mikä johtaa erilaisiin tietoaineistoihin. Lääketieteelliset tiedot sisältävät usein aikasarja-attribuutteja (mikä tarkoittaa, että tutkimusten ja laboratoriokokeiden päivämäärät ovat kliinisestä näkökulmasta erittäin tärkeitä), joten tutkijoiden on käsiteltävä näitä tietojoukkoja erityisesti ottaen aikatekijän huomioon. [18]

2. Koska sairaalatietojärjestelmät tai sairaalatietokannat on pääasiassa suunniteltu talous-/laskutustarkoituksiin eivätkä lääketieteellisiin/kliinisiin tarkoituksiin, voi olla erityisen haastavaa saada korkealaatuista dataa kliiniseen tietojenlouhintaan [20][21].

3. Yhdysvalloissa monet sairaalat eivät esimerkiksi käytä täysin sähköisiä potilastietojärjestelmiä. Näin ollen suuri osa lääketieteellisistä tiedoista (erityisesti laboratoriokokeiden tulokset) on paperimuodossa, mikä puolestaan johtaa siihen, että lääketieteelliset tiedot ovat usein elektronisen saatavuuden osalta puutteellisia [22]. Lisäksi suuri osa historiallisista potilastiedoista on paperimuodossa tai skannattuina digitaalisessa muodossa, joten niitä ei voi käyttää tietojenlouhintaan ilman merkittävää tietojen valmistelua. [18]

Terveydenhuollon tutkijoiden on varmistettava potilasrekisterin suojaus ja käsiteltävä potilastietoja HIPAA(Health Insurance Portability and Accountability Act)-sääntelyn mukaisesti. Onnistuneet luottamuksellisuutta säilyttävät strategiat sisältävät potilastietojen de-identifiointi- ja anonymisointistrategiat, jotka ovat tarpeen HIPAA:n ja muiden ihmistutkimusta koskevien säädösten noudattamiseksi. [18] Ter-

veydenhuollon sovelluksissa tietojenlouhinnassa on kuitenkin yhtä tärkeää varmistaa potilasturvallisuus ja ylläpitää herkkien tietojen turvallisuutta ja luottamuksellisuutta kuin tietojoukkojen saattaminen muiden tutkijoiden saataville [23]. Useissa tapauksissa tämä pysyy herkkänä tasapainona, kun tietojen laatu ja saatavuus vaikuttavat tarpeeseen suojata potilastietojen luottamuksellisuutta [24].

Terveystietojen käytössä on myös oikeudellisia näkökohtia. Esimerkiksi lääketieteellisen tiedon tiedonlouhinta voi paljastaa aiemmin tuntemattomia lääketieteellisiä virheitä, mikä voi puolestaan johtaa oikeustoimiin terveydenhuollon tarjoajia vastaan. Yksinkertaiset kirjoitusvirheet tai tiedon syöttövirheet voivat paljastua tietojen esikäsittelyn ja/tai tiedonlouhintaa koskevissa arvioinneissa ja tulkinnassa. [18] Nämä ongelmat voivat johtaa epäilyttävien kuvioiden löytämiseen lääketieteellisessä käytännössä. Tämä voi olla tärkeä näkökohta Yhdysvalloissa, missä lääketieteellisen virheen korvauslait ovat tiukkoja ja ihmiset ovat oikeustoimintakannalla [25]. Mahdollisuus lisääntyneeseen vastuuseen voi selittää, miksi terveydenhuollon tarjoajat epäröivät tarjota jopa de-identifioitua tai anonymisoitua lääketieteellistä tietoa tietojenlouhinnan tarkoituksiin [18].

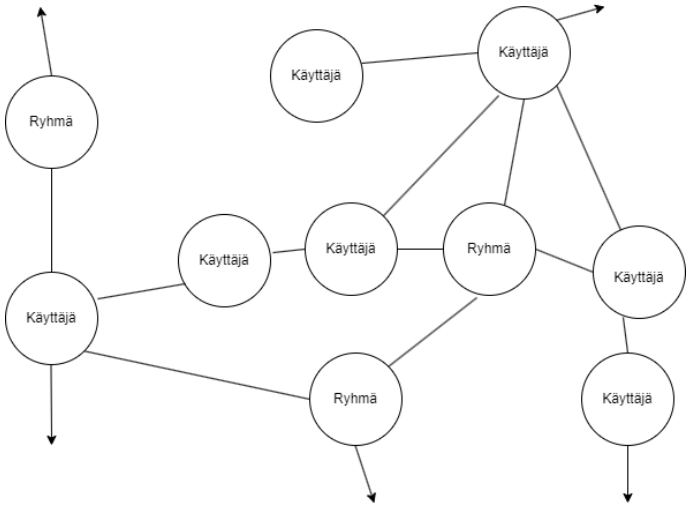
## 3.2 Tiedonlouhinta sosiaalisessa mediassa

Sosiaalisen median vauraus muuttaa vallankumouksellisesti arkeamme ja synnyttää samalla valtavan määrän tietoa. Suurten tietomäärien louhiminen ja analysointi sosiaalisessa mediassa antavat meille mahdollisuuden ymmärtää, miten ihmiset kommunikoivat, vuorovaikuttavat ja tekevät yhteistyötä verkossa. Lisäksi se on hyödyllistä suunnittelijoille ja yrityksille. Tämän avulla he voivat parantaa sosiaalisen median järjestelmiä ja alustoja paremmin vastaamaan ihmisten tarpeita. Näin ollen sosiaalisen median suurten tietomäärien louhiminen ja analysointi ovat olleet yksi kiinnostavimmista tutkimusalueista ja ne houkuttelevat yhä enemmän tutkimus yhteisön panostusta. [26][27][28]

Sosiaalisen median kasvua ohjaavat seuraavat haasteet: (1) Kuinka käyttäjä saadaan kuulluksi? (2) Minkä tiedonlähteen käyttäjä valitsee? (3) Kuinka käyttäjäkokemusta voidaan parantaa? Vastaukset näihin kysymyksiin piilevät sosiaalisen median tietojen syvyyksissä. [29]

Samoin kuin muissa sosiaalisen verkoston aineistoissa, yleistä on käyttää graafiesitystä sosiaalisen median aineistojen tutkimiseen. Graafi koostuu joukosta solmuja (node) ja kaaria (linkki). Yksilöt ovat tyypillisesti solmuja graafissa, ja yhteydet tai suhteet yksilöiden välillä (solmut) esitetään kaarina graafissa. Graafiesitys on luonteva sosiaalisen verkoston aineistoista peräisin oleville tiedoille, joissa yksilöt luovat verkoston ystävien, luokkatoverien tai liikekumppaneiden kanssa. [30]

Yhteisön tunnistamisen tarkoitus on löytää yhteisörakenne graafissa. Linkkianalyysialgoritmien soveltaminen sosiaalisen median tietoihin voi havaita ryhmiä, jotka eivät ole heti ilmeisiä [31]. Linkin ennustaminen tarkoittaa kykyä ennustaa, milloin uusia suhteita muodostuu, ja se tunnetaan linkin ennustamisen ongelmana [32].




Kuva 3.1: Hypoteettinen kaaviokuva tyypillisestä sosiaalisen verkostoitumisen sivuston grafiikkarakenteesta. Huomaa linkit käyttäjäsolmujen ja ryhmäsolmujen välillä. Nuolilla osoitetaan linkkejä suurempiin osiin graafista.
[30]

Sosiaalisen median tiedonlouhiminen tuo mukanaan monia haasteita, jotka vaativat huolellista huomiota ja innovatiivisia lähestymistapoja. Yksi keskeisistä haasteista on roskapostin tunnistaminen ja käsittely. [33][34] Sosiaalisen median alustoilla roskaposti voi olla merkittävä ongelma, ja sen erottaminen aidosta ja hyödyllisestä sisällöstä on tärkeää tiedonlouhimisen tehokkuuden kannalta. Lisäksi sosiaalisen median alakategorioiden käyttämien erilaisten formaattien moninaisuus asettaa haasteita tiedonlouhimisen automatisoinnille ja standardoinnille. Esimerkiksi kuvat, videot, tekstit ja linkit voivat kaikki olla osa sosiaalisen median sisältöä, mikä tekee tiedonlouhimisen monimutkaisemmaksi ja moniulotteisemmaksi tehtäväksi. [30]

Toinen merkittävä haaste liittyy sosiaalisen median sisällön jatkuvaan muutokseen ja rakenteen monimutkaisuuteen. Sosiaalisen median alustat ovat dynaamisia ympäristöjä, joissa sisältö ja käyttäjien vuorovaikutus muuttuvat jatkuvasti. Tämä vaatii tiedonlouhinnan menetelmien ja algoritmien jatkuvaa päivittämistä ja sopeutumista muuttuviin olosuhteisiin. Lisäksi sosiaalisen median verkostorakenteen monimutkaisuus asettaa haasteita yhteisörakenteiden tunnistamiselle ja linkkien ennustamiselle. Tiedonlouhinnan menetelmien on kyettävä käsittelemään suuria määriä dataa ja löytämään merkityksellisiä yhteyksiä ja kaavoja verkostojen monimutkaisesta rakenteesta. [35][30]

## 3.3 Koulutustietojen louhinta

E-oppimisresurssien, instrumentaalisten opetusohjelmien, Internetin käytön opetuksessa ja opiskelijatietokantojen perustaminen ovat luoneet suuria varastoja opetusdataa. Perinteiset oppilaitokset ovat käyttäneet monia vuosia tietojärjestelmiä, jotka tallentavat runsaasti mielenkiintoista tietoa. Nykyään verkkopohjaiset opetusjärjestelmät ovat kasvaneet eksponentiaalisesti ja johtaneet meidät tallentamaan valtavan määrän potentiaalista dataa monista eri lähteistä ja eri tarkkuustasoilla [36]. Uusien opiskeluympäristöjen, kuten sulautuvan oppimisen (blended learning BL),

virtuaali- ja tehosteympäristöjen, mobiili- ja ubiikin oppimisen sekä pelioppimisen, esiintyminen on saanut aikaan merkittävän määrän tietoa opiskelijoiden toiminnasta. Näissä ympäristöissä syntyy runsaasti opetuksellisesti arvokasta tietoa, joka on kuitenkin liian massiivista manuaaliseen analyysiin. Siksi tarvitaan automaattisia työkaluja tällaisen datan käsittelyyn. Tämä tietomäärä muodostaa arvokkaan resurssin opetusdatan muodossa, jonka analysoiminen ja ymmärtäminen on keskeistä opiskelijoiden oppimisprosessien syvällisessä tutkimuksessa ja hyödyntämisessä. [37]

Itse asiassa yksi suurimmista haasteista, joita oppilaitokset kohtaavat tänään, on opetusdatan eksponentiaalinen kasvu ja tämän datan muuntaminen uusiksi oivalluksiksi, jotka voivat hyödyttää opiskelijoita, opettajia ja hallintoa [38]. Kaksi erilaista yhteisöä ovat kehittyneet saman aiheen ympärille yhteisellä kiinnostuksella siitä, miten opetusdataa voidaan hyödyntää koulutuksen ja oppimistieteen hyväksi [39][37].

Koulutusdatan louhinta (Educational Data Mining, EDM) keskittyy kehittämään menetelmiä, jotka tutkivat kouluympäristöistä peräisin olevaa ainutlaatuista tyyppiä dataa [40]. Sitä voidaan myös määritellä sovellukseksi tiedonlouhintaan tälle erityiselle datatyypille, joka tulee kouluympäristöistä vastaamaan tärkeisiin koulutuskysymyksiin [41][37].

Oppimisanalytiikka (engl. Learning Analytics, LA) voidaan määritellä datan mittaamiseksi, keräämiseksi, analysoimiseksi ja raportoimiseksi oppijoista ja heidän konteksteistaan, oppimisen ymmärtämisen ja optimoinnin tarkoituksessa sekä oppimisympäristöissä että niiden ulkopuolella [42]. Tässä määritelmässä on kolme keskeistä elementtiä: data, analyysi ja toiminta [43][37].

Oppimisanalytiikka ja koulutustietojen louhinta tarjoavat mahdollisuuden ymmärtää oppilaiden yksilöllisiä oppimistarpeita ja tarpeita. Analysoimalla oppilaiden suoritusta ja käyttäytymistä voidaan tunnistaa tehokkaita opetusstrategioita ja tarjota yksilöllistä tukea oppimisprosessin eri vaiheissa. [44][45]

Koulutusympäristöissä kertyy runsaasti monimuotoista dataa eri lähteistä, kuten oppimisalustoista, kyselyistä ja oppimispäiväkirjoista. Tämä data sisältää sekä strukturoitua että epästrukturoitua tietoa, mikä tekee sen hallinnasta ja analysoinnista haastavaa. [15][37]

Koulutustietojen louhinta ja oppimisanalytiikka tarjoavat arvokasta tietoa koulutuspolitiikan suunnittelulle ja päätöksenteolle. Analysoimalla esimerkiksi oppimistuloksia ja opetuksen tehokkuutta voidaan tunnistaa alueita, joilla koulutusjärjestelmää voidaan parantaa ja resursseja kohdentaa tehokkaammin. [46][37]

Koska koulutustietojen louhinta koskettaa henkilökohtaista oppimistietoa ja käyttäytymistä, on tärkeää kiinnittää erityistä huomiota eettisiin kysymyksiin ja yksityisyydensuojaan. On varmistettava, että oppilaiden tiedot käsitellään ja säilytetään turvallisesti ja että heidän yksityisyyttään suojataan asianmukaisesti. [47][48]

# 4 Yhteenveto

Tiedonlouhinnan yleisimpiin menetelmiin kuuluvat klusterointi, luokittelu, poikkeamien havaitseminen ja yhdistäminen. Näiden menetelmien ja tekniikoiden avulla voidaan löytää arvokasta tietoa suurista tietomassoista. Tätä arvokasta tietoa voidaan käyttää useilla aloilla, kuten terveydenhuollossa, sosiaalisessa mediassa tai koulutuksessa.

Terveydenhuollossa tiedonlouhinnan menetelmistä käytetään poikkeamien havaitsemista, klusterointia ja luokittelua. Lääketieteellisen tiedonlouhinta voi paljastaa aiemmin tuntemattomia terveydenhuollossa tehtyjä virheitä ja auttaa oikaisemaan näitä. Tiedonlouhintaa voidaan myös käyttää erilaisten sairauksien ennustamiseen, sekä auttamaan lääkäreitä tekemään kliinisiä päätöksiä diagnoosien kanssa.

Tiedonlouhinta sosiaalisessa mediassa tarjoaa merkittäviä mahdollisuuksia ymmärtää käyttäjien käyttäytymistä, vuorovaikutusta ja yhteisörakenteita verkossa. Analysoimalla suuria tietomääriä voidaan tunnistaa käyttäjien käyttäytymismalleja ja vuorovaikutustapoja, mikä auttaa suunnittelijoita ja yrityksiä parantamaan sosiaalisen median järjestelmiä ja alustoja vastaamaan paremmin käyttäjien tarpeisiin.

Koulutuksessa tiedonlouhinta soveltuu moniin eri konteksteihin, kuten e-oppimisresurssien ja instrumentaalisten opetusohjelmien opetuksessa. Verkkopohjaiset opetusjärjestelmät ja uudet oppimisympäristöt, kuten sulautuva oppiminen ja mobiilioppiminen, tuottavat valtavan määrän dataa opiskelijoista. Tämä data tarjoaa arvokasta tietoa oppimisprosessista, mutta sen manuaalinen analysointi on

mahdotonta. Siksi tarvitaan työkaluja, kuten koulutustiedonlouhintaa, oppimisanalytiikkaa ja muita tiedonlouhintamenetelmiä, joiden avulla voidaan automaattisesti analysoida ja ymmärtää opetusdataa. Koulutustietojen louhinnalla ja oppimisanalytiikalla voidaan tunnistaa oppilaiden yksilöllisiä oppimistarpeita ja tarpeita sekä löytää tehokkaita opetusstrategioita. Analysoimalla oppilaiden suoritusta ja käyttäytymistä voidaan myös tehdä päätöksiä koulutuspolitiikan suunnittelusta ja resurssien kohdentamisesta.

Tiedonlouhinta tarjoaa laajan valikoiman haasteita ja mahdollisuuksia eri aloilla. Yleisesti ottaen, yksi suurimmista haasteista on suuren ja monimuotoisen tiedon hallinta ja analysointi. Tämä voi sisältää tiedon keräämisen, puhdistamisen, ja tiedon laadun varmistamisen. Lisäksi tietosuoja ja eettiset kysymykset, kuten yksityisyydensuoja ovat jatkuvasti kasvava huolenaihe. Erityisesti kun käsitellään herkkiä tietoja, kuten terveystietoja, asiakastietoja tai oppilastietoja.

Toisaalta, tiedonlouhinnalla on valtavasti potentiaalia tarjota arvokasta tietoa ja oivalluksia eri aloilla. Esimerkkejä mahdollisuuksista ovat sairauksien riskien ennustaminen terveydenhuollossa, käyttäjien käyttäytymisen ymmärtäminen ja trendien tunnistaminen sosiaalisessa mediassa, sekä oppimisstrategioiden optimointi ja koulutuspolitiikan suunnittelu koulutuksen alalla.

# Lähdeluettelo


[1] S. Sumathi ja S. Sivanandam, ”Introduction to data mining principles”, *Introduction to data mining and its applications*, s. 1–20, 2006.

[2] F. Coenen, ”Data mining: past, present and future”, *The Knowledge Engineering Review*, vol. 26, nro 1, s. 25–29, 2011.

[3] P. Smyth, ”Data mining: data analysis on a grand scale?”, *Statistical methods in medical research*, vol. 9, nro 4, s. 309–327, 2000.

[4] J. Rowley, ”The wisdom hierarchy: representations of the DIKW hierarchy”, *Journal of information science*, vol. 33, nro 2, s. 163–180, 2007.

[5] R. L. Ackoff, ”From data to wisdom”, *Journal of applied systems analysis*, vol. 16, nro 1, s. 3–9, 1989.

[6] C. A. Pushpam ja J. G. Jayanthi, ”Overview on data mining in social media”, *International Journal of Computer Sciences and Engineering*, vol. 5, nro 11, s. 147–157, 2017.

[7] U. Fayyad, G. Piatetsky-Shapiro ja P. Smyth, ”From data mining to knowledge discovery in databases”, *AI magazine*, vol. 17, nro 3, s. 37–37, 1996.

[8] V. Chandola, A. Banerjee ja V. Kumar, ”Anomaly detection: A survey”, vol. 41, nro 3, heinäkuu 2009, ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. url: https://doi.org/10.1145/1541880.1541882.

[9] P.-N. Tan, M. Steinbach ja V. Kumar, ”Data mining cluster analysis: basic concepts and algorithms”, *Introduction to data mining*, vol. 487, s. 533, 2013.

[10] Z. Birch, ”An efficient data clustering method for very large databases”, teoksessa *Proceedings of the 1996 ACM SIGMOD international conference on management of data (SIGMOD'96). ACM, New York*, 1996, s. 103–114.

[11] C. M. Bishop, ”Pattern recognition and machine learning”, *Springer google schola*, vol. 2, s. 645–678, 2006.

[12] E. W. Weisstein, ”Hypersphere”, *https://mathworld. wolfram. com/*, 2002.

[13] G. Kesavaraj ja S. Sukumaran, ”A study on classification techniques in data mining”, teoksessa *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, IEEE, 2013, s. 1–7.

[14] S. Agarwal, ”Data Mining: Data Mining Concepts and Techniques”, teoksessa *2013 International Conference on Machine Intelligence and Research Advancement*, 2013, s. 203–207. DOI: 10.1109/ICMIRA.2013.45.

[15] J.-J. Yang, J. Li, J. Mulder et al., ”Emerging information technologies for enhanced healthcare”, *Computers in industry*, vol. 69, s. 3–11, 2015.

[16] S. K. Sharma, N. Wickramasinghe ja J. N. Gupta, ”Knowledge management in healthcare”, teoksessa *Creating knowledge-based healthcare organizations*, IGI Global, 2005, s. 1–13.

[17] N. Jothi, N. A. Rashid ja W. Husain, ”Data Mining in Healthcare – A Review”, *Procedia Computer Science*, vol. 72, s. 306–313, 2015, The Third Information Systems International Conference 2015, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.12.145. url: https://www.sciencedirect.com/science/article/pii/S1877050915036066.

[18] I. Yoo, P. Alafaireet, M. Marinov et al., ”Data mining in healthcare and biomedicine: a survey of the literature”, *Journal of medical systems*, vol. 36, s. 2431–2448, 2012.

[19] R. Ichise ja M. Numao, ”Learning first-order rules to handle medical data”, *NII journal*, vol. 3, nro 2, s. 9–14, 2001.

[20] J. G. Jollis, M. Ancukiewicz, E. R. DeLong, D. B. Pryor, L. H. Muhlbaier ja D. B. Mark, ”Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research”, *Annals of internal medicine*, vol. 119, nro 8, s. 844–850, 1993.

[21] P. E. Dans, ”Looking for answers in all the wrong places”, *Annals of internal medicine*, vol. 119, nro 8, s. 855–857, 1993.

[22] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage ja W. E. Hammond, ”Medical data mining: knowledge discovery in a clinical data warehouse.”, teoksessa *Proceedings of the AMIA annual fall symposium*, American Medical Informatics Association, 1997, s. 101–105.

[23] J. J. Berman, ”Confidentiality issues for medical data miners”, *Artificial intelligence in medicine*, vol. 26, nro 1-2, s. 25–36, 2002.

[24] A. M. Berger ja C. R. Berger, ”Data mining as a tool for research and knowledge development in nursing”, *CIN: Computers, Informatics, Nursing*, vol. 22, nro 3, s. 123–131, 2004.

[25] K. J. Cios ja G. W. Moore, ”Uniqueness of medical data mining”, *Artificial intelligence in medicine*, vol. 26, nro 1-2, s. 1–24, 2002.

[26] X. Yang ja K. Li, ”Social media data mining and knowledge discovery under wireless network”, English, *Wireless Networks*, vol. 27, nro 5, s. 3375–3376, heinäkuu 2021, Copyright - © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021;

Last updated - 2024-02-09. url: https://www.proquest.com/scholarly-journals/social-media-data-mining-knowledge-discovery/docview/2548029942/se-2.

[27] R. W. Lariscy, E. J. Avery, K. D. Sweetser ja P. Howes, "Monitoring public opinion in cyberspace: How corporate public relations is facing the challenge", *Public Relations Journal*, vol. 3, nro 4, s. 1–17, 2009.

[28] S. Moro, P. Rita ja B. Vala, "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach", *Journal of Business Research*, vol. 69, nro 9, s. 3341–3351, 2016, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2016.02.010. url: https://www.sciencedirect.com/science/article/pii/S0148296316000813.

[29] P. Gundecha ja H. Liu, "Mining social media: a brief introduction", *New directions in informatics, optimization, logistics, and production*, s. 1–17, 2012.

[30] G. Barbier ja H. Liu, "Data Mining in Social Media", teoksessa *Social Network Data Analytics*, C. C. Aggarwal, toim. Boston, MA: Springer US, 2011, s. 327–352, ISBN: 978-1-4419-8462-3. DOI: 10.1007/978-1-4419-8462-3_12. url: https://doi.org/10.1007/978-1-4419-8462-3_12.

[31] I. King, J. Li ja K. T. Chan, "A brief survey of computational approaches in social computing", teoksessa *2009 International Joint Conference on Neural Networks*, IEEE, 2009, s. 1625–1632.

[32] D. Liben-Nowell ja J. Kleinberg, "The link prediction problem for social networks", teoksessa *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, s. 556–559.

[33] N. Agarwal ja H. Liu, *Modeling and data mining in blogosphere*. Morgan & Claypool Publishers, 2009.

[34] A. Java, P. Kolari, T. Finin, T. Oates et al., "Modeling the spread of influence on the blogosphere", *UMBC TR-CS-06-03*, 2006.

[35] G. Lakshmanan ja M. Oberhofer, "Knowledge discovery in the blogosphere: Approaches and challenges", *IEEE internet computing*, vol. 14, nro 2, s. 24–32, 2010.

[36] C. Romero ja S. Ventura, "Educational data science in massive open online courses", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, nro 1, e1187, 2017.

[37] C. Romero ja S. Ventura, "Educational data mining and learning analytics: An updated survey", *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 10, nro 3, e1355, 2020.

[38] R. S. Baker, "Big data and education", *New York: Teachers College, Columbia University*, 2015.

[39] R. S. J. de Baker ja P. S. Inventado, "Chapter X: Educational Data Mining and Learning Analytics", *Comput. Sci*, vol. 7, s. 1–16, 2014.

[40] B. Bakhshinategh, O. R. Zaiane, S. ElAtia ja D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years", *Education and Information Technologies*, vol. 23, s. 537–553, 2018.

[41] C. Romero, M.-I. López, J.-M. Luna ja S. Ventura, "Predicting students' final performance from participation in on-line discussion forums", *Computers & Education*, vol. 68, s. 458–472, 2013.

[42] C. Lang, G. Siemens, A. Wise ja D. Gasevic, *Handbook of learning analytics*. SOLAR, Society for Learning Analytics ja Research New York, 2017.

[43] G. Siemens, "Learning analytics: The emergence of a discipline", *American Behavioral Scientist*, vol. 57, nro 10, s. 1380–1400, 2013.

[44] A. Pardo ja G. Siemens, ”Ethical and privacy principles for learning analytics”, *British journal of educational technology*, vol. 45, nro 3, s. 438–450, 2014.

[45] R. S. Baker, K. Yacef et al., ”The state of educational data mining in 2009: A review and future visions”, *Journal of educational data mining*, vol. 1, nro 1, s. 3–17, 2009.

[46] S. Slade ja P. Prinsloo, ”Learning analytics: Ethical issues and dilemmas”, *American Behavioral Scientist*, vol. 57, nro 10, s. 1510–1529, 2013.

[47] S. Dawson, D. Gašević, G. Siemens ja S. Joksimovic, ”Current state and future trends: A citation network analysis of the learning analytics field”, teoksessa *Proceedings of the fourth international conference on learning analytics and knowledge*, 2014, s. 231–240.

[48] R. Ferguson, ”Learning analytics: drivers, developments and challenges”, *International Journal of Technology Enhanced Learning*, vol. 4, nro 5-6, s. 304–317, 2012.