Haaga-Helia
ammattikorkeakoulu

HUOM! Tämä on alkuperäisen artikkelin rinnakkaistallenne. Rinnakkaistallenne saattaa erota alkuperäisestä sivutukseltaan ja painoasultaan.

Käytä viittauksessa alkuperäistä lähdettä:

Kauttonen, J., Haukkala, M. & Lahtinen, A. (21.03.2023) Laadukas data tekoälyn moottorina. eSignals PRO. http://urn.fi/URN:NBN:fi-fe2023032132675

PLEASE NOTE! This in an electronic self-archived version of the original article. This reprint may differ from the original in pagination and typographic detail.

Please cite the original version:

Kauttonen, J., Haukkala, M. & Lahtinen, A. (21.03.2023) Laadukas data tekoälyn moottorina. eSignals PRO. http://urn.fi/URN:NBN:fi-fe2023032132675

Copyright: © 2023 by the authors and Haaga-Helia University of Applied Sciences. Licensed under the terms and conditions of the Creative Commons Attribution (CC BY NC SA) license (https://creativecommons.org/licenses/by-nc-sa/4.0/).

Laadukas data tekoälyn moottorina

Janne Kauttonen, Mikko Haukkala & Anna Lahtinen

Tekoälysovellukset vaativat laadukasta dataa tuottaakseen laadukkaan lopputuloksen. Datan laatua voidaan arvioida pilkkomalla datan laatu eri ulottuvuuksiin, jolloin laadun arviointi helpottuu. Suurin hyöty laadukkaan datan tunnistamisessa syntyy kuitenkin kontekstiymmärryksestä.

Datan laatu on tekoälyn käyttöönoton suurimpia haasteita

Tutkimuksen mukaan on useita datan laadun tekijöitä, jotka vaikuttavat tekoälysovellusten toimintaan. Asiaankuuluvuus ja datan täydellisyys nousevat usein esiin selvitettäessä haastavimpia datan laadun ulottuvuuksia. Liiketoimintaymmärrys on tärkeä kerättävää dataa suunniteltaessa ja määriteltäessä. Keskeisenä haasteena nähdään mm. vaikeus ennakoida, millaista dataa tulevia liiketoimintatarpeita varten tulisi kerätä. Tässä yhteydessä liiketoiminnan ja teknisen toteutuksen ja/tai IT:n keskustelu ja vuorovaikutus ovat ratkaisevia (Haukkala, 2022.)

Tekoälyn hyödyt ovat kiistattomat. Sen tuomat mahdollisuudet uuden liiketoiminnan kehittämisessä ja ongelmien ratkaisemisessa on havaittu monella alalla. Aiemmat tutkimukset kuitenkin osoittavat, että datan laatu on yksi suurimmista tekoälyn käyttöönoton haasteista. Microsoftin (2018) suomalaisyrityksille tekemän kyselyn mukaan 80 % yrityksistä koki, että käytetty data ei ollut tarpeeksi valmista soveltamiseksi käytännön tekoälyratkaisuihin.

Kun liiketoimintatarpeet on kirkastettu, yritysten kannattaa tarvittavan datan keräämiseksi laatia systemaattinen suunnitelma. Tekoälyn käyttötapauksen varsinainen työstäminen alkaa yrityksen datavarantojen kartoittamisesta. Tämä tarkoittaa käytännössä sitä, että selvitetään millaista dataa yritys käyttää ja tuottaa toiminnassaan, missä tämä data sijaitsee ja onko se omassa hallussa, paljonko sitä on ja miltä osin data vastaa sitä, mitä käyttötapaus tarvitsee. Datan käsittelyyn ja jalostamiseen täytyy varata riittävästi aikaa, sillä datan laatu tulee määrittelemään lopputuloksen.

Datan kartoitusprosessi käytännön esimerkein

Datan kartoituksessa kokonaisuus kannattaa pilkkoa pienempiin osiin, joka helpottaa huomattavasti kartoitustyötä. Erityisesti seuraavat ohjaavat kysymykset voivat auttaa tätä prossia (muokattu Alanko-Turusesta ym. 2022):

- Mitä dataa meillä on? Listataan kaikki yrityksen keräämä ja tallentama data joka liittyy käyttötapauksiin.
- Millä tavoin data liittyy liiketoimintatarpeisiin ja -tavoitteisiin?
- Missä dataa varastoidaan? Datan fyysinen tallennuspaikka ja tallennusformaatti.
- Kuka datan omistaa? Onko data juridisesti yrityksen omaisuutta vai jonkun muun (esim. asiakkaan).
- Saammeko datan hallintaamme tai käyttöömme? Datasta ei ole hyötyä ellei sitä saada käyttöön ja hyödynnettäväksi myös kaupallisiin tarkoituksiin.

- Onko dataa riittävästi? Onko datan määrä riittävä tavoitteisiin nähden.
- Onko datan laatu hyvä vai onko siinä puutteita? Datan suuri määrä ei koskaan korvaa laatua.
- Onko data hyödynnettävässä muodossa? Data on muunnettava ja jalostettava oikeaan muotoon ennen hyödyntämistä ja voi vaatia huomattavia resursseja.
- Mitä datalähteitä puuttuu ja miten data kerätään? Tunnistetaan puutteet ja tehdään suunnitelma keräämiselle.

Seuraavassa avaamme näitä kysymyksiä käytännössä esimerkkien avulla. Otamme kohteeksi kolme kuvitteellista PK-yritystä, jotka haluaisivat kehittää liiketoimintaansa datan ja data-analytiikan, mm. tekoälyn, avulla. Käymme läpi yritysten tilannetta edellä mainittujen apukysymysten kautta ja tunnistamme keskeiset datahaasteet, sekä hahmottelemme toimintasuunnitelman siitä, miten yritysten kannattaisi edetä. Esimerkit mukailevat aitoja tapauksia, joita olemme kohdanneet projektin puitteissa.

Esimerkki 1: Työtarpeen ennakointi hoiva-alan yrityksessä

Yritys tarjoaa kotihoidon ja avustajien palveluita koteihin. Yrityksen tavoite on optimoida työvuoroja ja ennakoida työvoimatarvetta lähiviikoille. Yrityksen käytössä on kattavasti omaa dataa sisältäen mm. asiakasrekisterin, hoitajarekisterin ja hoitosopimukset. Dataa on useiden vuosien ajalta ja se on kokonaan omassa omistuksessa.

Tällä hetkellä data on kuitenkin yrittäjän omalla työkäytössä olevalla kannettavalla tietokoneella erillisissä tiedostoissa, erillisessä pilvipalvelussa (esim. kalenterit), sekä osittain myös paperimuodossa mapeissa (esim. sopimukset). Lisäksi kaikkea dataa ei ole kerätty ja muotoiltu samalla tavalla ja myös ajankohdat, jolloin data tallennetaan, vaihtelee suuresti. Esimerkiksi työntekijöiden ja asiakkaiden tiedot ovat sekalaisessa muodossa, koska niitä ei ole kerätty samanlaiselle pohjalle.

Tunnistetut datahaasteet: Data on hajallaan eri paikoissa. Mapeissa oleva data ei ole hyödynnettävässä muodossa ennen digitaaliseen muotoon saattamista. Dataa ei ole kerätty yhdenmukaisesti ja sen laatu vaihtelee riippuen siitä kuka työntekijä dataa on kirjannut ja milloin. Datassa on puuttuvia tietoja. Henkilötietojen tallennus ja säilytys ei noudata säännöksiä esim. GDPR:n osalta.

Toimintasuunnitelma: Tehdään yhteinen suunnitelma datan keräykselle siten, että se yhdenmukaistetaan koko yrityksen tasolla. Tiedot kerätään ja tallennetaan aina samalla tavalla johdonmukaisesti, mahdollisimman nopeasti ja samassa formaatissa (esim. sama lomake kaikilla ja se täytetään kokonaan). Siirrytään kokonaan digitaalisiin asiakirjoihin ja tietokantoihin, jotka toimivat tietoturvallisessa pilvipalvelussa. Hankitaan tähän liittyvät IT-palvelut erilliseltä palvelutoimittajalta kilpailuttamalla. Lähdetään pohtimaan tarkemmin datan hyödyntämistä ja analyysejä vasta näiden työvaiheiden jälkeen.

Esimerkki 2: Toimitusketjun optimointi logistiikka-alalla

Yritys vastaanottaa, varastoi ja kuljettaa tavaroita asiakkaille. Yrityksen tavoitteena on kulujen karsiminen (esim. kuljetusmatkat ja polttoaine) muun muassa minimoimalla tavaroiden käsittely ja säilytysajat, sekä toimitusten ajomatkat. Yrityksen käytössä on omaa dataa sisältäen mm. asiakasrekisterin, rahtikirjat, varastokirjat ja kalustorekisterin. Tämän lisäksi yritys on tunnistanut avoimen karttapalvelun, jota se on kiinnostunut hyödyntämään omassa toiminnassaan.

Yritys on hankkinut IT-palvelut alihankkijan kautta ja kaikki yrityksen data sijaitsee heidän palvelimellaan, jonne data siirtyy ERP-sovelluksen kautta. Dataa on tallennettu kattavasti useiden vuosien ajan, mutta toisaalta rahtikirjat ovat edelleen paperisessa muodossa tai sekalaisina valokuvina kansiossa.

Tunnistetut datahaasteet: Yrityksellä ei ole suoraa hallintaa tietokantaan, vaan se on tällä hetkellä alihankkijan hallinnassa. Tietokannassa olevan datan laatu on epäselvä, koska sitä ei ole tarkastettu pitkään aikaan. Osa mahdollisesti hyödyllisestä datasta on paperimuodossa tai valokuvamuodossa, joten tämä data ei ole sellaisenaan suoraan hyödynnettävissä. Avoimen karttapalvelun datan kaupallisen hyödyntämisen ehdot ovat epäselvät.

Toimintasuunnitelma: Selvitetään aluksi IT-alihankkijan kanssa tietokantaan pääsy ja datan nykytilanne. Selvitetään miten yksittäisten tavaroiden koko elinkaaren (vastaanotto, varastointi ja kuljetus) tiedot saadaan yhdistettyä eri tietokannoista. Tutkitaan millainen datan laatu on, esimerkiksi onko siinä huomattavia aukkoja tai epätarkkuuksia. Rahtikirjojen osalta luodaan käytännöt datan digitaalisesta tallentamisesta, esimerkiksi skannaamalla kaikki paperikopiot, jolloin ne ovat tasalaatuisia. Selvitetään karttapalvelun osalta käyttöehdot ja tehdään tarvittaessa sopimus palvelutoimittajan kanssa kaupallisesta käytöstä.

Esimerkki 3: Asiakastuntemuksen lisääminen ja tuotekehitys komponenttivalmistuksessa

Yritys suunnittelee, valmistaa ja myy elektronisia komponentteja suoraan asiakkaille. Yritys haluaisi lisätä myyntiä ja kehittää tuotteitaan hyödyntäen asiakastietoja ja komponentteja hyödyntävien asiakkaiden laitteiden keräämää lokitietoa. Yrityksellä on runsaasti dataa asiakkaista ja myynnistä, lisäksi asiakkaiden komponentteja hyödyntäviin laitteisiin kertyy erilaista lokidataa komponenttien toiminnasta.

Data on pääosin laadukasta, mutta lokidatoissa on aukkoja ja dataa on kerätty vasta 3kk ajan. Asiakasdataa on 16v ajalta, eli sitä on riittävästi. Datan muoto ei kuitenkaan ole yhdenmukaista koko ajalta, vaan datan keräystapa on vaihtunut ainakin kolme kertaa vuosien aikana ohjelmistopäivitysten myötä.

Tunnistetut datahaasteet: Lokitiedot ovat asiakkaiden omistuksessa ja yrityksellä ei ole niihin suoraa pääsyä. Lokidataa on liian vähän kattavan analyysin tekemiseen. Asiakasdataa on paljon, mutta data ei ole suoraan vertailukelpoista eri vuosien välillä.

Toimintasuunnitelma: Selvitetään miten asiakasdata on tarkalleen muuttunut vuosien varrella ja korjataan, sekä täydennetään dataa tarvittaessa. Jos data todetaan tämän jälkeen riittävän hyväksi, voidaan edetä pilottivaiheeseen, jossa lähdetään kokeilemaan erilaisia data-analyysimenetelmiä (mm. tunnistamalla asiakasryhmiä). Lokidatan osalta neuvotellaan asiakkaiden kanssa datan hyödyntämisestä ja saamiseksi yrityksen omaan tutkimuskäyttöön, tarvittaessa korvausta vastaan. Arvioidaan miten paljon dataa puuttuu, ovatko puutteet vakavia ja miten ne vaikuttavat analyyseihin. Mikäli puutteet eivät estä datan hyödyntämistä, kerätään dataa kunnes määrä on riittävä analyysien aloittamiseen.

Tekoälyn hyödyntämistä edeltää dataan liittyvien ongelmien ratkaisut

Kun datahaasteet on selvitetty toimintasuunnitelmien mukaisesti, tekoälyn käyttötapauksen varsinainen toteutus voisi tapahtua esimerkiksi hyödyntäen optimointia (esimerkit 1 ja 2) ja ennustavia koneoppimismalleja liittyen aikasarja-analyysiin (esimerkit 1 ja 3), sekä asiakkaiden ryhmittelyä klusterointianalyysiä käyttäen (esimerkki 3). Emme kuitenkaan mene näihin menetelmiin syvemmälle tässä kirjoituksessa.

Analyysimenetelmä ja käytetty algoritmi ratkaisee kuitenkin hyvin pitkälle sen, millaista dataa toteuttaminen vaatii ja mihin muotoon data on saatettava. Tässä yhteydessä erityisesti IT ja data-analyysin osaajien tuominen mukaan keskusteluihin on tärkeää. Mikäli riittävää dataosaamista ei löydy yrityksen sisältä, mikä on tyypillistä PK-yrityksissä, sitä voidaan hankkia paitsi rekrytoinneilla myös ostopalveluna. Dataan erikoistuneita konsultteja ja IT-palveluiden tarjoajia löytyy runsaasti ja niitä kannattaa hyödyntää.

Nykyään data-alustaksi valitaan useimmiten pilvipalvelu, joka mahdollistaa myös edistyneet tekoälyratkaisut yhdestä ja samasta palvelusta. Suurimmat alan toimijat ovat AWS, Google Cloud ja MS Azure. Ennen varsinaisten tekoäly- ja analytiikkaratkaisun toteuttamista, on yrityksen datavarannot ja datastrategia saatava ensin riittävälle tasolle ja kaikki dataongelmat ratkaistua. Tekoälyn hyödyntäminen voi alkaa vasta sen jälkeen.

Kirjoitus on osa Tekoälyinnovaatioekosysteemillä kilpailukykyä pk-yrityksille (AI-TIE) -hanketta. Euroopan aluekehitysrahaston ja Uudenmaan liiton tuella AI-TIE -hanke edistää PK-yritysten liiketoiminnan kehittämistä ja kasvua tekoälyratkaisuja hyödyntäen osana Euroopan unionin covid-19-pandemian johdosta toteuttamia toimia. Hankkeen toteuttavat Haaga-Helia ja Laurea ammattikorkeakoulut yhdessä tiiviin partneriverkoston kanssa. Lisätietoa AI TIE:stä: www.aistories.fi

Lähteet

Alanko-Turunen, M., San Miguel, E., Kauttonen, J., Ruohonen, A., Humala, I., Lagstedt, A. 2022. AI in Business – Tekoäly liiketoiminnassa [Verkkokurssi]. AI-TIE, Tekoälyinnovaatioekosysteemillä kilpailukykyä PK-yrityksille.

Haukkala, M. 2022. Data Quality in Artificial Intelligence. Opinnäytetyö. Talous, hallinto ja markkinointi (YAMK) International Business Management tutkinto-ohjelma. Helsinki: Haaga-Helia ammattikorkeakoulu.

Microsoft News Center. 2018. Artificial Intelligence in Europe – Finland.