Osaamista
ja oivallusta
tulevaisuuden
tekemiseen

Lasse Huotari

Ammattilaisen oppimisalusta koneoppimisen ja datatieteen tehokkaaseen opiskeluun

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tuotantotalous

Insinöörityö

25.10.2019

Tekijä Otsikko Sivumäärä Aika	Lasse Huotari Ammattilaisen oppimisalusta koneoppimisen ja datatieteen tehokkaaseen opiskeluun 46 sivua 25.10.2019
Tutkinto	insinööri (AMK)
Tutkinto-ohjelma	tuotantotalous
Ammatillinen pääaine	international ICT business
Ohjaajat	Lehtori Anna Sperryn Cloud Techinical & Solutioning Leader Hemmo Hiltunen

Insinöörityön aiheena oli työn ohessa tapahtuva datatieteen ja koneoppimisen koulutuksen tehostaminen. Työn tavoite oli rakentaa validoitu ammattilaisen interaktiivinen oppimisalusta, jossa yhdistetään yleisesti käytössä olevat prosessimallit sekä käytännönharjoitukset.

Opinnäytetyö toteutettiin IT-alan asiakasyritykselle. Datatieteen ja koneoppimisen nopea kehitys vaikeuttaa asiakasyrityksen kykyä tarjota kattavaa ja ajantasaista koulutusta toimihenkilöille.

Asiakasyrityksen nykytilaa ja käytössä olevia koulutusmenetelmiä selvitettiin yrityksen datatieteen parissa työskentelevien toimihenkilöiden haastatteluiden avulla. Haastatteluissa selvitettiin toimihenkilöiden omia kokemuksia nykyisin tarjolla olevista koulutuksista sekä kartoitettiin heidän toiveitaan ja näkemyksiään uuden oppimisalustan vaatimuksista ja toiminnallisuuksista.

Tutkimuksessa selvisi, että Cross-Industry Standard Process for Data-Mining on yleisesti käytetty datatieteen prosessimalli. Sen soveltuvuutta oppimisalustan pohjaksi testattiin syksyllä 2018 Metropolia ammattikorkeakoulussa järjestetyssä liiketoiminnan analytiikka -kurssilla.

Opinnäytetyössä kehitettiin interaktiivinen oppimisalusta prosessimallia ja käytännönharjoituksia yhdistäen, joka validoitiin testikäytöillä asiakasyrityksessä. Testikäytöistä kerätty palaute oli positiivista, ja osoitti että pienellä jatkokehityksellä ehdotettu oppimisalusta on tehokas keino järjestää datatieteen koulutusta.

Avainsanat	Datatiede, Koulutus, Oppimisalusta

Author	Lasse Huotari
Title	A Professional Learning Platform for Effective Learning in Machine Learning and Data Science
Number of Pages	46 pages
Date	25 October 2019
Degree	Bachelor of Engineering
Degree Programme	Industrial Management
Professional Major	International ICT business
Instructors	Anna Sperryn, Senior Lecturer Hemmo Hiltunen, Cloud Technical & Solutioning Leader

The purpose of this thesis was to explore professional data science and machine learning education in work life. The main objective was to create a validated interactive learning platform for professionals which combines the process models that are used in data science today and hands-on practices.

The case organization is a major IT company. Data science and machine learning are study fields which are developing fast, and this makes it hard for the case organization to offer up-to-date, comprehensive education for employees.

For this study, employees involved in data science were interviewed to gather feedback on how their organization has organized the current data science and machine learning education. In the same interviews, the employees were asked to give their ideas on how education should be arranged and what requirements and wishes they have for a learning platform.

The study and the research showed that Cross-Industry Standard Process for Data-Mining is a generally used data science process model. The suitability of CRISP-DM to be a base for a learning platform was tested in the fall of 2018 on the business analytics course which was held at Metropolia UAS.

Based on the results of this study an interactive learning platform was developed. After the developing phase, the platform was validated and tested by the employees of the case organization. The feedback from the validation was positive and showed that with small improvements the learning platform is an effective way to arrange data science education.

Keywords	Data Science, Education, Learning Platform

Sisällys

1 Johdanto 1

2 Menetelmät ja materiaalit 4

2.1 Tutkimussuunnitelma 4
2.2 Tietolähteet ja analyysimetodit 5

3 Nykytila-analyysi 6

3.1 Omaehtoinen kouluttautuminen 6
3.1.1 Internetportaalit 7
3.1.2 Omaehtoiset seminaarit ja koulutustilaisuudet 8
3.2 Asiakasyrityksen sisäinen koulutus 8
3.3 Datatieteilijän työskentelytavat ja prosessit yrityksessä 9
3.4 Nykytilan yhteenveto 11

4 Parhaat käytänteet oppimisalustaan 13

4.1 Datatiede 13
4.2 Tiedonlouhinta 15
4.3 Cross-Industry Standard Process for Data-Mining 16
4.3.1 Liiketoiminnan ymmärtäminen 17
4.3.2 Datan ymmärtäminen 19
4.3.3 Datan valmistelu 22
4.3.4 Mallintaminen 23
4.3.5 Arviointi 25
4.3.6 Tuotteistaminen 27
4.4 Chatbot 29
4.5 Git-versionhallintajärjestelmä 30
4.6 Oppimisalusta 31
4.7 Teoreettinen viitekehys 32

5 Ehdotuksen rakentaminen asiakasyritykselle 33

5.1 Katsaus ehdotuksen rakenteesta ja validoinnista 33
5.2 Kurssi 33

5.3 Interaktiivinen oppimisalusta 35
5.3.1 Alustan osa-alueet 35
5.3.2 Tietosisältö 35
5.3.3 Tehtävät 35
5.3.4 Kommunikaatio 40
5.3.5 Alustan tekninen toteutus 40
6 Ehdotuksen validointi 41
6.1 Yhteenveto 41
6.2 Oppimisalustan koekäytön palaute 42
6.3 Opinnäytetyön arviointi 44
Lähteet 46

1 Johdanto

Opinnäytetyö käsittelee työn ohessa tapahtuvan opiskelun parhaiden käytänteiden selvittämistä ja niihin perustuvan käytännönläheisen opintoalustan toteuttamista kohdeyrityksessä. Opinnäytetyössä keskityttiin erityisesti koneoppimisen ja datatieteiden osa-alueisiin. Tämä tutkimus on tehty globaalin IT-yrityksen konserniin kuuluvalle suomalaiselle tytäryhtiölle.

Yritys on kiinnostunut kehittämään omaa koulutustaan datatieteen ja koneoppimisen osalta, koska koulutuksen tehostamisella on suoria hyötyjä heidän liiketoimintansa tehokkuuteen ja siten kannattavuuteen. Koulutuksen kehittämisellä pystytään nopeuttamaan uuden työntekijän sopeutumista rooliinsa ja lisäämään valmiutta omaksua uusia tehtäviä.

Tällä hetkellä yrityksessä käytetään vaihtelevia menetelmiä, ja jokaisen työntekijän vastuulla on selvittää itse hänelle sopivat tavat opetella uusia aiheita. Opetusmateriaalin paljouden ja sen monimuotoisuuden takia täysi itseopiskelu ei tue uuden tehtävän ja menetelmien omaksumista tarpeeksi. Näin ollen parhaisiin käytänteisiin perustuva oppimisalusta ja sen faktaperusteinen dokumentointi sekä avaaminen koetaan tärkeäksi.

Tutkimuksessa käytetty data on peräisin haastatteluista yrityksen työntekijöiltä sekä toimihenkilöiltä, jotka ovat aiheita opettaneet korkeakoulutasolla. Dataa kerättiin myös Metropolia Ammattikorkeakoulussa syksyllä 2018 pidetyllä Liiketoiminnan analytiikkakurssilla palautteen muodossa. Kurssilla pilotoitiin myös tämän tutkimuksen teoriaosuuden tietosisältöä käytännöntasolla erilaisten tehtävien muodossa.

Teoriaosuus koostuu pääosin Cross-Industry Standard Process for Data-Mining -prosessimallin ja avainasemassa olevien teknologioiden kirjallisuuteen ja niitä käsitteleviin eri artikkeleihin.

Tutkimuksessa käytettävät käsitteet ja teoria pohjautuvat alan kirjallisuuteen, valmiisiin harjoituksiin sekä pedagogisiin menetelmiin.

Tutkimuksen pääkohteena on datatieteen ja koneoppimisen koulutuksen kehittäminen. Niiden käsitteet määritellään seuraavasti:

> Datatieteen avulla tietomassasoista – suurista ja pienistä – voidaan löytää riippuvuuksia ja säännönmukaisuuksia. Sen avulla voidaan luoda selitettäviä ja ennustavia malleja – toisin sanoen ymmärrystä menneestä, nykytilasta ja tulevaisuudesta. Tuloksia hyödynnetään tietoon perustuvan päätöksenteon työkaluna, prosessien optimoinnissa sekä toimintojen automatisoinnissa. [1.]

Koneoppimisessa on kyse tietokoneen ohjaamisesta oppimaan esimerkeistä ja sen avulla voidaan luoda ennustavia malleja, jotka pystyvät ratkaisemaan niille asetettuja ongelmia [2].

Tutkimuksessa käytetyllä käsitteellä koulutus tarkoitetaan tapaa parantaa työntekijän tietotaitoa kyseessä olevasta asiakokonaisuudesta, ja hänen valmiuttaan hyödyntää tietoa osana toimenkuvaansa. Koulutuksen osana toimivat harjoitukset, joilla havainnollistetaan käytännön tasolla kyseessä olevan asian toiminnallisuus ja pyritään luomaan työntekijälle käytännön osaaminen kyseessä olevasta aihekokonaisuudesta. Toisena osa-alueena on harjoituksien lomassa tarjottava teoriatieto, joka avaa kunkin aiheen taustoja ja tarkoitusta työntekijälle. Näin autetaan rakentamaan tarvittavan teoriaosaamisen aihealueesta.

Parhaiden käytäntöjen teoria nojautuu vahvasti CRISP-DM:ään. CRISP-DM on edelleen yksi käytetyimpiä data mining -prosessimalleja, ja sen toiminnallisuus sekä sovellettavuus luo erinomainen pohjan suunnitellulle oppimisalustalle. [3.]

Yritys on IT-alan yritys, joka on perustettu 1900-luvun alkupuolella Yhdysvalloissa. Yrityksen palveluksessa työskentelee 200-400 tuhatta työntekijää, ja sen liikevaihto vuonna 2018 oli 50-100 miljardia dollaria. Pääliiketoiminta-alueet yrityksellä ovat pilviratkaisut, tekoäly ja kognitiivisen koneoppimisen ratkaisut sekä erilaiset software- ja hardware-konsultoinnin palvelut.

Asiakasyritys on osa globaalin emoyhtiön konsernia. Se on perustettu 1900-luvun alussa, ja sen liikevaihto on vuonna 2017 ollut 100-250 miljoonaa euroa. Se työllisti vuonna 2017 noin 700 henkilöä. Asiakasyrityksen toimialat ovat samat kuin konsernin.

Tämän tutkimuksen tavoitteena on luoda nykyaikainen oppimisalusta koneoppimisen ja datatieteen koulutukseen. Ensin tutkitaan asiakasyrityksen sisäisiä mahdollisuuksia henkilökunnan kouluttamiseen, toiseksi nykyisin käytössä olevan datatiedeprojektien CRISP-DM-mallin soveltumista oppimisalustan tietosisällön pohjaksi.

Tavoitteiden saavuttamiseksi tutkimus pyrkii vastaamaan seuraaviin kysymyksiin:

- Minkä tyyppiset harjoitukset ammattilaiset kokevat tehokkaimmiksi?
- Onko CRISP-DM relevantti malli käytettäväksi modernin oppimisalustan pohjana?
- Miten datatieteilijä voidaan kouluttaa tehokkaammin toteuttamaan projekteja yhtenäisen prosessin periaatteiden mukaan ja näin saada jatkumoa työhönsä?
- Minkälainen on nykyinen tyypillinen kouluttautumisprosessi yrityksen työntekijöillä?
- Miten yritys voisi kehittää sisäistä koulutustaan paremmaksi?

Tutkimus keskittyy vain datatieteiden ja koneoppimisen harjoituksiin ja koulutukseen, vaikkakin ne edustavat vain pientä osiota yrityksen liiketoiminnasta. Näiden osa-alueiden osalta kehitys on valtavan nopeaa, ja siksi yritys kokee niiden koulutuksen tehokkuuden parantamisen erityisen tärkeäksi. Lisäksi koko yrityksen liiketoiminnan kattavan koulutuksen tutkiminen olisi epärelevanttia työn laadukkuuden ja tavoitteiden kannalta.

Tutkimuksen tuotos on validoitu ehdotus interaktiivisesta oppimisalustasta, joka käyttää moderneja teknologioita hyväkseen tarjotakseen tehokkaampaa koulutusta työntekijöille. Oppimisportaalin avulla datatieteilijä saavuttaa nopeammin työn vaatiman osaamistason sekä oppii käyttämään oppeja käytännössä. Tehokkaampi opiskelu tuottaa yritykselle liiketoiminnan hyötyä lisäämällä työtehoa sekä lyhentämällä uuden roolin omaksuntaan käytettyä aikaa. Näin ollen työntekijä pääsee nopeammin tuottavan työn pariin.

Tutkimus koostuu kuudesta osasta, ja sen rakenne on seuraava: Osa 1 on tutkimuksen esittely. Osa 2 käsittelee metodeja ja materiaaleja, joita raportissa on käytetty. Osa 3 on nykytila-analyysi ja kuvaa, kuinka työntekijät kokevat yrityksen tarjoaman koulutuksen tällä hetkellä. Osa 4 kuvaa nykyisiä parhaita käytänteitä datatieteilijöiden viitekehyksissä,

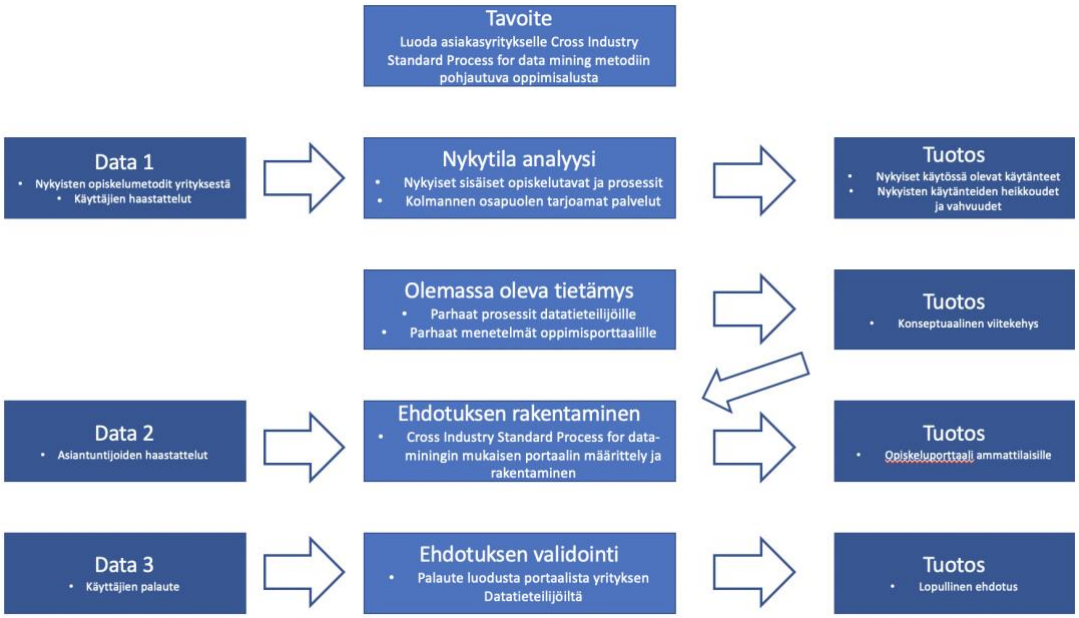
oppimisalustan rakennetta sekä interaktiivisen oppimisalustan luomiseksi tarvittavia moderneja teknologioita. Osa 5 käsittää alustavan ehdotuksen uudesta koulutusmenetelmästä ja osa 6 sisältää kyseisen ehdotuksen validoinnin.

2 Menetelmät ja materiaalit

Tässä osiossa käsitellään tutkimuksen metodologiaa, tutkimussuunnitelmaa, dataa sekä käytettyjä menetelmiä, joiden tarkoituksena on tehostaa datatieteilijöiden koulutusta.

2.1 Tutkimussuunnitelma

Tutkimuksen tutkimussuunnitelma on kuvattu kuvassa 1.:



Kuva 1. Tutkimuksen tutkimussuunnitelma

Kuten ylläolevasta kuvasta nähdään, tutkimus alkaa tavoitteen määrittelyllä. Tämän jälkeen havainnoidaan nykytilaa, kartoitetaan nykyistä tietoa, haastatellaan asiantuntijoita ja tehdään ehdotelma interaktiivisesta oppimisalustasta.

2.2 Tietolähteet ja analyysimetodit

Tutkimuksessa käytetty data on kerätty useista eri tietolähteistä. Se sisältää haastatteluja ja palautteita työntekijöiltä ja opiskelijoilta sekä tutkimuksen tekijän havaintoja kurssin sisällöstä ja onnistumisesta. Tietolähteet on kuvattu alla.

Haastattelut ja tapaamiset

Haastattelut toimivat pääasiallisena tietolähteenä tutkimuksen ehdotuksen rakentamisessa ja sen kohdentamisessa asiakasyritykselle. Haastatteluihin osallistui yrityksen työntekijöitä analytiikan osastolta, joilla on kokemusta itse datatieteestä sekä sen opettamisesta ja kouluttamisesta eri organisaatioissa.

Haastatteluissa kysyttiin asiantuntijoiden havaitsemia keinoja tehokkaaseen opiskeluun, sekä heidän mielipidettänsä asiakasyrityksen nykytilasta sekä toiveita, minkälaisia kouluttautumisvaihtoehtoja olisi ideaalitilanteessa käytettävissä. Haastatteluista kerättiin data nauhoitteina.

Taulukko 1. Haastattelujen ja tapaamisten data

Datatyyppi	**Osallistujat**	**Päivämäärä**	**Aihe**	**Dokumentaatio**
Haastattelu 1	Lasse Huotari	29.8.2018	Nykytila ja Analytiikan opetus	Nauhoite 1
	Työntekijä 1, Tekninen asiantuntija			
Haastattelu 2	Lasse Huotari	29.8.2018	Nykytila ja Analytiikan opetus	Nauhoite 2
	Työntekijä 2, Arkkitehti			
Haastattelu 3	Lasse Huotari	24.9.2018	Uudet menetelmät analytiikan opettamiseen	Nauhoite 3
	Työntekijä 2, Arkkitehti			
Haastattelu 4	Lasse Huotari	21.11.2018	Datatieteen nykytila asiakasyrityksessä	Nauhoite 4
	Työntekijä 3, Data Scientist			
	Työntekijä 4, Data Scientist			
	Työntekijä 5, Data Scientist			

Kuten taulukosta 1 nähdään, nykytila-analyysin data kerättiin henkilöhaastatteluista ja niiden nauhoitteista. Haastattelut suoritettiin kasvotusten.

Kurssilta saadut palautteet

Nykyisten datatiedeprojektin viitekehyksen parhaiden käytänteiden soveltuminen testattiin Metropoliassa pidetyllä liiketoiminnan analytiikan kurssilla. Kurssista kerättiin muistiinpanoja, ja palautteita. Kurssipalautteita kerättiin kurssin lopussa opiskelijoilta, ja niitä tuli yhteensä 14 kappaletta. Palautteet annettiin aikavälillä 26.11.2018 – 13.12.2018. Palautteet olivat opiskelijoiden vapaasti kirjoittamia ajatuksia kurssin sisällöstä, ja sen mielekkyydestä omiin opintoihin nähden. Lisäksi opiskelijat ottivat kantaa palautteissaan kurssin rakenteeseen ja sen mielekkyyteen, ja asiasisällön haastavuuteen.

3 Nykytila-analyysi

Tämä osio käsittelee asiakasyrityksen nykytilaa ja paneutuu tarkemmin koulutuksessa käytettyihin metodeihin. Osion ensimmäinen osa käsittelee omaehtoisen kouluttautumisen menetelmiä sekä vapaasti saatavilla olevaa koulutusaineistoa. Toinen osa käsittelee yrityksen järjestämää koulutusta. Kolmas osa käsittelee datatieteilijän työskentelyä yrityksessä ja heidän käyttämiään prosessimalleja.

3.1 Omaehtoinen kouluttautuminen

Datatieteilijöiden työ on hyvin nopeasti uudistuvaa, ja uusia menetelmiä syntyy todella nopeasti. Näin ollen omaehtoinen kouluttautuminen on erityisen tärkeässä roolissa. Omaehtoisella koulutuksella saavutetaan jatkuvaa kehittymistä, mutta se yleisesti keskittyy enemmän työkaluihin kuin itse prosessiin tai projektin toteutukseen. Asiakasyritys tukee työntekijöiden omaehtoista kouluttautumista.

3.1.1 Internetportaalit

Datatieteilijöillä on käytössä useita eri koulutusalustoja internetissä, joista osa on vapaasti saatavilla ja osa maksullisia. Internetportaaleissa koulutus tapahtuu yleensä harjoitustehtävien kautta, mutta teoriaosaamista tuetaan joko tekstimuodossa tai videotallenteilla.

Internetportaalit voidaan jakaa luonnollisesti kahteen osioon. Niihin, joissa opetus on yliopistomaista kurssien suorittamista, ja niihin, joissa kurssien kesto on huomattavasti lyhyempi, ja kurssin voi suorittaa jopa muutamassa tunnissa.

Yleisesti pidempiä yliopistomaisia kursseja tarjoavat yliopistot. Esimerkiksi Harvardilla ja MIT:lla (Massachusetts Institute of Technology) on kattavat luentosarjat, joihin voi osallistua. Yliopistojen lisäksi on muutamia toimijoita, jotka tarjoavat pitkiä luentosarjoja. Näistä esimerkkinä toimii Coursera, joka tuottaa luentoja yhteistyössä eri yliopistojen kanssa.

Toinen internetportaalien muoto ovat niin sanotut pikakurssit, jotka tyypillisesti kestävät muutaman tunnin. Näillä kursseilla paneudutaan teoriaan hyvin lyhyesti, ja tarkoituksena on ennemminkin nopeasti kuvata aiheeseen liittyvät avainkäsitteet. Päätyökalu näillä kursseilla on kuitenkin itse nopeissa harjoitustehtävissä, joiden avulla datatieteilijä voi löytää nopeita vastauksia ongelmiinsa, tai saada alkeet uudesta metodista, ja jatkaa tämän jälkeen opiskelua muuten.

Pikakursseista esimerkkinä on Cognitive Class, joka tarjoaa virtuaalisia ilmaiskursseja kaikille asiasta kiinnostuneille. Pikakursseja tarjoavat portaalit eivät yleensä ole virallisia opetuslaitoksia, joten he eivät myönnä todistuksia tai tutkintoja.

Datatieteilijät jakavat usein tekemiään harjoituksia vapaasti internetissä, ja näitä hyödynnetään valtavasti opetellessa uusia menetelmiä. Julkaisualustana toimivat yleensä blogit tai Github, joten harjoitusten löytäminen saattaa usein viedä aikaa ja olla työlästä.

Tutkimuksen haastattelujen perusteella internetportaaleja pidetään varsin hyvinä uuden asian ylätasolla opiskelemiseen, mutta syvemmälle mentäessä ja asian vaikeutuessa

lähiopetus koetaan tarpeelliseksi. Internetportaalien isoimpina haasteina koetaan materiaalin jälkikäteisen tutkimisen vaikeus. Portaaleissa suurin osa koulutuksesta tapahtuu videoiden välityksellä, ja jälkikäteen tiettyyn asiaan nopeasti palaaminen videotallenteesta on hankalaa. Myöskään videon litterointia ei koeta hyvänä vaihtoehtona. Osasta koulutusmateriaalia on saatavilla muistiinpanojen tapaiset jälkeismateriaalit, joissa on viittaus videon kohtaan, jossa asia löytyy tarkemmin selitettynä. Tämä koetaan hyväksi tavaksi jakaa muistiinpanot kurssista, mutta tapa on vähiten käytetty.

3.1.2 Omaehtoiset seminaarit ja koulutustilaisuudet

Toinen tapa kouluttaa omaehtoisesti itseään on osallistua erilaisiin seminaareihin ja koulutustilaisuuksiin. IT-yhteisössä järjestetään usein koulutustapahtumia, joissa pääsee kuulemaan kollegoilta, miten he tekevät työtään ja minkälaisten asioiden kanssa he työskentelevät.

Seminaareja ja koulutustilaisuuksia järjestävät sekä yritykset että yhteisöt. Yritys on ollut vuonna 2018 mukana järjestämässä esimerkiksi Mimmit koodaa -nimistä tapahtumasarjaa, jossa alasta kiinnostuneita naisia koulutetaan datatieteen ja tekoälyn osa-alueilla.

Vertaiskoulutus on haastattelujen perusteella havaittu toimivaksi tavaksi työskennellä, mutta sen vapaaehtoisuuden vuoksi jatkuvan kouluttautumisen varmentaminen on usein hyvin hankalaa.

3.2 Asiakasyrityksen sisäinen koulutus

Asiakasyritys edellyttää työntekijöitään kouluttautumista omaan työtehtäväänsä. Se järjestää koulutusta uuden työroolin alkaessa. Kurssit ovat joko lähiopetusta, tai itsenäisiä pakollisia suorituksia.

Datatieteilijöille yritys järjestää esimerkiksi omia koulutusohjelmiaan, joiden kesto on noin 50 - 80h. Koulutusohjelmaan osallistuu kerrallaan kymmeniä datatieteilijöitä, jolloin kollegoiden kanssa ideoiden vaihtaminen ja tiedon siirtäminen helpottuu. Lisäksi

yrityksessä järjestetään erilaisia Bootcampeja, eli hieman lyhyempiä koulutuksia, joissa käsitellään otsikkotasolle valittua state of the art -metodia tai menetelmää.

Asiakasyrityksellä on erilaisia koulutusohjelmia, joissa koulutetaan työntekijöitä yrityksen tarpeisiin. Esimerkkinä on vastavalmistuneiden opiskelijoiden -ohjelma, joka on kaksi vuotta kestävä koulutus. Koulutuksen sisältö määräytyy kyseisen työroolin mukaan, ja se on tarkoitettu juuri valmistuneille tai valmistumassa oleville henkilöille.

3.3 Datatieteilijän työskentelytavat ja prosessit yrityksessä

Datatieteilijät työskentelevät varsin itsenäisesti, ja toimenkuvan asiantuntijuusluonteen vuoksi käytettävät työkalut ja prosessit ovat hyvin vaihtelevia. Yritys ei ole asettanut datatieteilijöilleen pakollista prosessia, miten projekteja tulee tehdä, mutta yrityksen sisällä on datatieteilijöille vakiintuneita käytänteitä, minkä mukaan datatiedeprojekteja tehdään. Tämä osio kuvaa yleisellä tasolla näitä vakiintuneita käytänteitä.

Perusolettamuksella datatieteilijän vastuille kuuluu liiketoiminnan ymmärtäminen, datan käsittely, datan mallintaminen ja mallien tuotteistaminen. Tässä osiossa käsitellään datatieteilijän työkaluja kuhunkin edellä kuvattuun vastuuseen nähden sekä tarvittavaa prosessinäkökulmaa. Käytännössä datatieteilijät käyttävät itse valitsemaansa työkalua näihin tehtäviin, mutta asiakkaan vaatimukset saattavat ohjata päätöstä.

Liiketoiminnan ymmärtämiseen ei tarvita ulkoisia työkaluja. Yleisesti hyväksi havaittu metodi on järjestää niin sanottu service design -workshop, jossa asiakkaan kanssa käydään läpi heidän liiketoimintaansa, alkavan projektin tavoitteita ja tärkeimpiä mittareita tavoitteiden saavuttamisen varmistamiseksi.

Service design tarkoittaa palvelumuotoilua, jonka alkuperäisenä määritelmänä on käyttäjälähtöinen palveluiden suunnittelu muotoilun menetelmin. Liiketoiminnan ymmärtämisen vaiheessa palvelumuotoilulla kuitenkin tarkoitetaan projektin palvelumuotoilua, jossa muotoillaan konsultointipalvelu asiakasyritykselle.

Service design -workshopin tavoitteena on tuottaa sekä projektin asiakasyritykselle, että datatieteilijälle käsitys siitä, minkälaisia asioita projektissa koitetaan tavoitella ja mitkä ovat oikeat menetelmät tavoitteiden saavuttamiseksi. Workshopissa on datatieteilijän kanssa paikalla asiakasyrityksen liiketoiminnan ymmärtävät henkilöt, sekä mahdollisesti asiakasyrityksen analyytikot, datantieteilijät sekä tietokanta-asiantuntijat, jos heitä on.

Workshopissa muodostuneen palvelumuotoilusuunnitelman avulla johdetaan projektia ja siinä kuvataan tarkasti muun muassa käytetty data, miten se saadaan käyttöön ja onko olemassa esimerkiksi joitakin rajoitteita, miten dataa voidaan analysoida. Esimerkiksi, jos se sisältää asiakastietoja, asiakasyritys voi rajoittaa datan käyttöä, henkilöitä, jotka voivat osallistua datan prosessointiin tai datan varastointia.

Liiketoiminnan ymmärtämisen jälkeen datatieteilijä rupeaa työskentelemään itse datan parissa, ja hänen ensimmäinen tehtävänsä onkin saada valittu data valittuun työkaluun. Yleisesti tässä vaiheessa käytetään apuna erilaisia tietokantoja ja niiden yhdistämismenetelmiä. Työn tekee monesti asiakasyrityksen tietokantaosaaja, mutta joissakin tapauksissa datatieteilijän tulee tukea tietokantaosaajaa esimerkiksi ETL-prosessin (extract, transform, load -prosessi) kanssa.

Datan siirron jälkeen alkaa datatieteilijän työn monimuotoisemmat osat. Nykytilassa prosessi on todella vaihteleva, ja kukin datatieteilijä voi vapaasti valita työkalunsa ja toimintatapansa itse. Lähtökohtaisesti datan ymmärrys antaa reunaehdot datatieteilijälle käytettävistä työkaluista, ja ne on sovitettava asiakkaan asettamiin ehtoihin.

Datan käsittelyyn datatieteilijä valitsee sopivan työkalun ja menetelmän ongelman luonteen mukaan. Yleisinä käytäntöinä voidaan pitää yrityksessä joko SPSS-tuoteperheeseen kuuluvan SPSS Modelerin käyttöä, tai avoimen lähdekoodin työkaluja. Ohjelmointikielenä avoimen lähdekoodin työkaluissa toimivat joko R, Python tai Scala, josta R ja Python ovat ehdottomasti suosituimpia.

Datan prosessoinnin aikana datatieteilijä valmistelee datan sopivaksi, varmistaa sen laadun sekä tekee analyyttisiä havaintoja datasta, joilla koitetaan selittää eri tapahtumia ja havaintoja. Lisäksi datatieteilijä tuottaa näistä asiakkaalle raportit valitsemallaan menetelmällä.

Analyysit tarkistetaan asiakkaan kanssa projektin väliarvioinnissa, joissa tarkennetaan service design -workshopissa asetettuja tavoitteita. Tavoitteet voivat muuttua workshopin tavoitteista, koska data voi tuottaa kiinnostavia havaintoja tai se osoittautuu kelpaamattomaksi alkuperäisiin tavoitteisiin.

Väliarvioinnista saadun hyväksynnän jälkeen datatieteilijä aloittaa itse mallinnustyöskentelyn, jossa tavoitteena on luoda asiakkaan kanssa sovitut mallit, joita päästään hyödyntämään käytännön sovellutuksissa. Mallit voivat olla koneoppivia malleja, syväoppivia malleja tai joissakin tapauksissa myös esimerkiksi loogisia päättelyitä tai sääntökoneita. Tämä riippuu projektin tavoitteesta ja järkevimmästä toteutustavasta.

Projektin viimeisessä vaiheessa tuotetaan tehdyistä malleista joko raportti tai tuotteistettu palvelu tai tuote. Tuotteistuksen tekee yleensä joku muu organisaatiosta kuin datatieteilijä, mutta datatieteilijä avustaa tuotteistuksesta vastaavaa henkilöä ymmärtämään mallien toimintalogiikkaa ja tuottaa hänelle tarvittavat dokumentaatiot malleista. Jos asiakkaalle tehdään vain raportti saavutetuista tuloksista, niiden tekeminen ja esittely on yleensä datatieteilijän vastuulla.

Asiakasyritykset haluavat yleensä siirtää osaamista myös omaan organisaatioon, joten datatieteilijän vastuulla on kouluttaa tarvittavat henkilöt käyttämään tuotteita ja tukea heitä luomiensa mallien päivittämisessä ja ymmärtämisessä.

Jos projektissa tavoitellaan pidempikestoista palvelua, datatieteilijä myös vastaa tuottamiensa mallien monitoroinnista, eli niiden toiminnan laadun ja tilan tarkkailuun soveltuvan toiminnallisuuden luomisesta. Hän rakentaa niin sanotut palauteloopit, joiden avulla malli voidaan kouluttaa tarvittaessa uudestaan.

3.4 Nykytilan yhteenveto

Vapaus valita työkalut ja prosessit helpottavat datatieteilijöiden työtä, ja he pystyvät nopeammin adaptoitumaan asiakkaan vaatimuksiin, mutta tietyn rakenteen puuttuminen

hankaloittaa kouluttautumista, koska ei ole olemassa yhtä selkeää tapaa tehdä työtä. Yhtenäisen koulutuksen ja työtapojen puute johtaa joissakin tapauksissa yhteentörmäykseen datatieteilijöiden välillä, ja tiedon siirtäminen toiselta toiselle hidastuu. Uuden työntekijän voi olla vaikea päästä mukaan kirjoittamattomaan vakiintuneeseen prosessiin, ja työn tehokkuus saattaa näin ollen laskea, vaikka uusi työntekijä olisikin työhön muuten pätevä.

Kurssitarjonta on hajanaista, eikä se ole aina johdonmukaista, minkä työntekijät kokevat haastavaksi. Työntekijät kaipaavat yhtenäisempää prosessia oppia yrityksen työkalut sekä yhteisen tavan tehdä töitä. Myös prosessien opettelemisen merkitys koetaan tärkeäksi, ja sen yhtenäistäminen sekä käytännöntason liittäminen erilaisiin työkaluihin ja projektin vaiheisiin esimerkkien kautta mielekkääksi lähestymistavaksi. Datatieteilijät kokevat tärkeäksi saada ymmärrystä myös tuotantotason toteutuksesta, koska välttämättä tuotantotason toteutuksessa ei voida käyttää välttämättä samoja menetelmiä, mitä on käytetty projektin ensivaiheissa hyväksi. Datatieteilijöiden toiveena on saada alusta, jossa yhdistyvät sekä yrityksen omat teknologiat, niiden tehokas läpikäyminen ja esimerkkien kautta opettaminen sekä prosessien havainnollistaminen.

Vahvuudet	Heikkoudet
Yritys tarjoaa monipuolisen ja runsaan kurssitarjonnan.	Kurssit ovat hajanaisia eikä niitä ole välttämättä kytketty toisiinsa.
Työ on itseohjautuvaa, ja työntekijä saa useasti asiakkaan kanssa päättää projektiin parhaiten soveltuvat käytännöt.	Datatiede kehittyy nopeasti, ja ison organisaation on hankala vastata aina yhtä nopeasti uusiin muutoksiin.
Organisaatio tarjoaa laajan vertaisryhmän, josta löytyy osaamista hyvin laajalla sektorilla.	Nykyisten työkalujen, ja projektien prosessimallien yhteys ei ole verkkokursseilla tarpeeksi

	johdonmukaista, eikä niitä käsitellä yhdessä.

4 Parhaat käytänteet oppimisalustaan

Tämä osio kuvaa parhaita käytänteitä datatieteen opiskeluun, prosessiin ja menetelmiin. Se keskittyy datatieteen opiskeluprosessin parantamiseen, mutta tarjoaa teorian eri menetelmien, kuten CRISP-DM:n osalta. Osiossa käydään lisäksi läpi menetelmiä, joilla itse opiskelun sisältöä voidaan paremmin konkretisoida opiskelijalle, sekä tuoda työn eri osa-alueita paremmin esille. Asiakasyrityksen nykytila-analyysin pohjalta tehdyt havainnot vahvistavat tarvetta yhtenäiselle, keskitetylle koulutusalustalle, jossa samaan aikaan tuetaan niin asiaosaamista kuin tuoteosaamista. Tutkimukseen valittu CRISP-DM on standardoitunut menetelmä datatiedeprojekteissa, vaikka käyttäjät eivät usein tietoisesti valitsekaan kyseistä metodia projektinsa prosessimalliksi. Lisäksi nykytila-analyysin pohjalta voidaan tehdä päätelmä, jossa interaktiivisten lähestymistapojen lisääminen koulutukseen on toimihenkilöiden mukaan havaittu tehokkaaksi tavaksi oppia uusia toimintatapoja ja menetelmiä. Käytössä olevien resurssien määrä rajoittaa täysin lähitoteutuksena toteutetun opetuksen mahdollisuutta, joten koulutuskokonaisuus on toteutettava modernina oppimisportaalina.

4.1 Datatiede

Datatiede on statistiikasta kehittynyt tieteen muoto, joka yhdistelee perinteisten statististen menetelmien lisäksi erilaisia konsepteja ja viitekehyksiä, kuten tekoälyä, koneoppimista ja IoT-ratkaisuja (Internet Of Things) [4]. Sen päätarkoituksena pelkistettynä on löytää mitä tahansa merkityksellistä tietoa datasta. Datatiedettä voidaan harjoittaa kahdella tasolla, joko tieteen ja tutkimuksen datatiedettä tai liiketoiminnan datatiedettä. Erona näillä kahdella datatieteen osa-alueella on niiden tapa käsitellä esimerkiksi koneoppimisen konsepteja ja käytänteitä siinä, missä tutkimusperäinen datatiede pyrkii luomaan uusia konsepteja ja käytänteitä niin liiketoiminnan datatiede

pyrkii pääsääntöisesti hyödyntämään näitä tutkimuksen jo kehittämiä konsepteja erilaisten valmiiden viitekehyksien avulla, esimerkiksi avoimeen lähdekoodin mallinnuskirjastoja hyödyntämällä. [5.]

Datatiedettä harjoittavia ammattilaisia kutsutaan datatieteilijöiksi. Datatieteilijöiden ammattikuvaan kuuluu useiden koodikielten hallitseminen, ohjelmistojen arkkitehtuurin ymmärtäminen sekä tietenkin tilastollisten menetelmien tunteminen. Datatieteilijöiden tehtävänä on rakentaa ohjelmistokokonaisuuksien osia, jotka käyttävät hyväksi aikaisemmin kuvattuja datatieteilijöiden osaamisalueen asioita kuin itse datatieteen ominaispiirteitä. [6.]

Datatiede keskittyy nimensä mukaisesti datapohjaiseen tieteen harjoittamiseen, jossa ideana on saada havaintoja käsiteltävästä datasta, ja käyttää näitä havaintoja tuomaan liiketoiminnalle hyötyä. Vaikka datatieteeseen varsinkin liiketoiminnan puolella liitetään vahvasti usein tekoäly ja koneoppiminen ei datatieteilijän itsetarkoitus ei ole tuottaa koneoppivia malleja, vaan ratkaista monialaisia ongelmia datavetoisesti, ja tuottaa kuhunkin ongelmaan paras ratkaisu. Niinpä datatietelijää ja datatiedettä voidaankin ajatella datavetoisena ratkaisusoveltamisena.

Datatieteen lähtökohtana on data, ja sen saatavuus on oletusarvona sille, että datatieteilijä voi tehdä työtänsä. Usein datatieteilijän toimenkuvaan liitetään virheellisesti myös datainsinöörin tehtävät, joihin lukeutuu muun muassa tietokantojen ylläpito sekä erilaisten datalakejen, eli yrityksen laaja tietovarasto, jossa kaikkea dataa pidetään varastossa, ennen kuin se haetaan sieltä käyttöön. [7.] Datatieteessä käsiteltävä data voi olla rakenteellista tai rakenteetonta. Rakenteeton data voi koostua kuvista, videoista, numeroista, tekstistä tai äänestä [8].

Rakenteellisella datalla viitataan rakenteellisessa tietokannassa säilytettävään dataan, kuten esimerkiksi SQL-tietokannan numeeriseen dataan. Se koostuu riveistä ja sarakkeista, ja data esiintyy aina samassa muodossa, se on samankaltaista ja se varastoidaan aina samoin. Rakenteeton data taas on esimerkiksi kuvia tai tekstiä, joita voi olla eri määriä, se voi sisältää erimäärän kirjaimia tai pikseleitä, ja sen asettaminen tiettyyn esimerkiksi taulukolliseen muotoon on mahdotonta. Se varastoidaan yksittäisinä tiedostoina ja kukin tiedosto esittää yhtä esimerkkitapausta. [8.] Datatieteilijä joutuu käsittelemään työssään molempia datan tyyppejä, ja niihin erikoistuneita metodeja.

Rakenteettoman datan lisääntymisen myötä datatieteilijöiden työ on jakaantunut vielä kahteen tutkimusosa-alueeseen akateemisen ja liiketoiminnan datatieteen lisäksi, rakenteellisen ja rakenteettoman datan tutkimusosa-alueeseen.

Datatiede mainittiin ensimmäisiä kertoja vuonna 1974 Peter Naurin kirjassa *Concise Survey of Computer Methods,* jossa hän julkaisi oman määritelmänsä silloin uudelle konseptille:

> "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences [9]."

Datatiede on kasvattanut siitä asti merkitystään maailmassa sekä liiketoiminnassa. Datatiedettä hyödynnetään perinteisen liiketoiminnan saralla, eri teollisuusaloilla, sairaanhoidossa, lääketieteessä, finanssisektorilla sekä julkisessa hallinnossa. Datatiede ulottuukin ennemmin tai myöhemmin kaikille eri toimialoille, ja sen merkitys kasvaa entisestään yritysten välisessä kilpailussa sekä sisäisessä onnistumisessa. [10.]

4.2 Tiedonlouhinta

Tiedonlouhinta on prosessikokoelma erilaisia työkaluja ja prosesseja, joilla yritykset kääntävät heidän tai muiden toimijoiden raakadatan hyödylliseen ja informatiiviseen muotoon. Tiedonlouhinnan perusperiaatteisiin kuuluu toistuvuuksien ja muuttujien riippuvuuksien löytäminen suuresta datamassasta. Näiden riippuvuuksien ja toistuvuuksien avulla voidaan tulkita paremmin yrityksen omaa liiketoimintaa tai sen markkinassa tai ympärillä tapahtuvia muutoksia ja syy-seuraus-suhteita. [11]

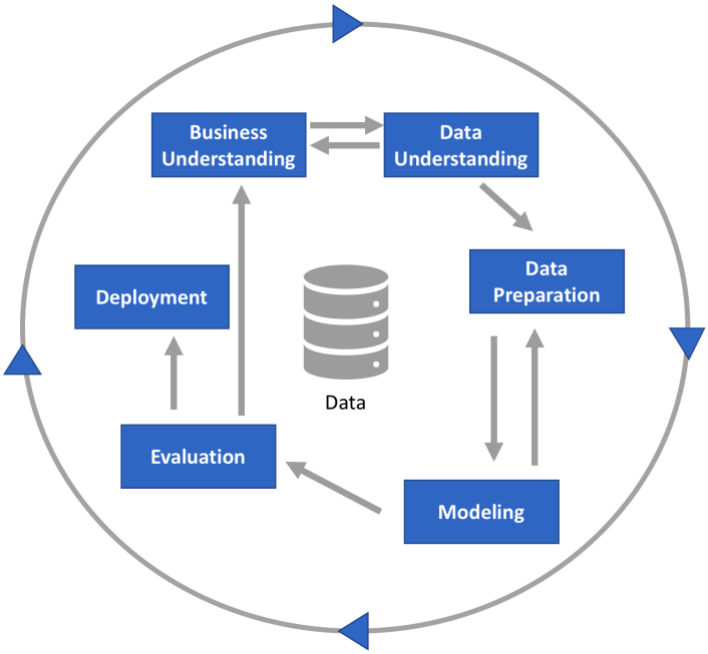
Tiedonlouhinta hyödyntää niin sanottua alhaalta ylös tekniikkaa, jossa ei esitetä ensin hypoteesia, jota koitetaan selittää datalla, vaan dataa tutkitaan statistisilla menetelmillä ja saadut havainnot muodostavat hypoteesit [11].

Tiedonlouhinta jakautuu viiteen eri vaiheeseen: datan keräämiseen ja sen tallentamiseen tietovarastoon, tietovaraston hoitamiseen ja säilyttämiseen, datan saatavuuteen ja järjestämiseen, sen prosessointiin ja lopuksi tulosten esittämiseen [11].

4.3 Cross-Industry Standard Process for Data-Mining

Cross-Industry Standard Process for Data-Mining (CRISP-DM) on laajasti käytetty poikkialainen standardiprosessi tiedonlouhintaan. Sillä on kaksi ominaisuutta, sitä voidaan käyttää metodologiana tai prosessimallina, metodologiana käytettäessä se kuvailee tyypilliset projektin vaiheet ja niihin liittyvät tehtävät, sekä selittää eri vaiheiden suhteen toisiin vaiheisiin. [12.]

Tämän osion teoreettinen tausta perustuu Smart Vision Europen [13] CRISP-DM-mallin mukaisiin vaiheisiin ja tehtäviin. Prosessimallina se tarjoaa standardisoidun työkalun projektin suunnittelemiseen ja dokumentoitiin. Kehyksen tarkoitus on olla iteratiivinen sekvensseihin perustuva kehys, joka sisältää kaikki data mining -osuudet. Tarpeen mukaan kehyksen sisällä voidaan liikkua vapaasti, myös taaksepäin, jos prosessin edellisiin vaiheisiin täytyy tehdä muutoksia. CRISP-DM koostuu kuudesta kohdasta, jotka on kuvattu kuvassa 2.



Kuva 2. CRISP-DM-kaavio. [3]

Eri palikoita yhdistävät nuolet osoittavat ideaalia etenemisjärjestystä, ja ulkoringissä olevat nuolet osoittavat mallin toistettavuutta, iteraatiota. Malli koostuu kuudesta

vaiheesta: liiketoiminnan ymmärtämisestä, datan ymmärtämisestä, datan valmistelusta, mallinnuksesta, arvioinnista sekä tuotteistamisesta. Keskiössä on itse data.

Seuraavissa osioissa on kuvattu tarkasti CRISP-DM:n eri vaiheet ja niistä muodostettavat tuotokset. Rakenteellisesti seuraava osio sisältää ensin tehtävän tai vaiheen kuvauksen, jonka jälkeen jokaisen tuotoksen sisältö on kuvattu.

4.3.1 Liiketoiminnan ymmärtäminen

CRISP-DM:n ensimmäinen vaihe on liiketoiminnan ymmärtäminen. Liiketoiminnan ymmärtämisen vaihe suoritetaan ennen kuin dataa tai työkaluja valitaan. Sen ideana on määrittää, mitä ja miksi projektilla halutaan tavoittaa. Vaihe sisältää neljä primitiivistä tehtävää, joista jokainen voi sisältää useamman alatehtävän. Neljä primitiivistä tehtävää ovat seuraavat:

- liiketoiminnan tavoitteiden määritely
- tilanteen määrittely
- tiedonlouhinnan tavoitteiden määrittely
- projektisuunnitelman tuottaminen.

Liiketoiminnan tavoitteiden määrittelyssä luodaan ymmärrys siitä, mitä yrityksen liiketoiminta haluaa saavuttaa tiedonlouhinnalla. Ensimmäinen asia on selvittää, mikä on varsinainen ongelma, joka halutaan ratkaista. Toinen kysymys on määrittää selkeästi, mitkä ovat liiketoiminnan tavoitteet, jotka saavutetaan, jos kyseinen ongelma saadaan ratkaistua. Kolmas selvitettävä on erilaiset rajoitteet, kuten projektille asetetun määräajan määritteleminen sekä kuinka projekti tulee suorittaa. Neljäs selvitettävistä asioista on projektin vaikutukset. Vaikutuksilla tarkoitetaan asiassa yhteydessä sitä, miten ongelmaan kehitetty ratkaisu tulee sopimaan yrityksen nykyisen liiketoiminnan kanssa.

Liiketoiminnan tavoitteiden määrittelystä tuotetaan kolme raporttia:

Projektin tausta - Raportissa kuvataan liiketoiminnan tilanne, joka ohjaa projektia. Oleellisena kysymyksenä raportissa on, miksi tämä projekti koetaan tarpeelliseksi ja minkä takia haluamme sen toteuttaa.

Liiketoiminnan tavoite –raportti, jossa kuvataan liiketoiminnan projektille asettamat tavoitteet. Raportti pyrkii vastaamaan kysymykseen siitä, mitä projektilla tavoitellaan ja miten. Raportin asettamat tavoitteet ovat tarkkoja, mutta laajoja.

Liiketoiminnan onnistumisen kriteerit, jossa kuvataan kaikki ne onnistumisen kriteerit, joilla liiketoiminta katsoo projektin onnistuneeksi. Asetetaanko esimerkiksi liiketoiminnasta saatavia tuloksia vertailuun ennen projektia tehtyihin havaintoihin nähden? Näin selvitetään, onko projekti onnistunut.

Tilanteen määrittelyssä tutkitaan yrityksen nykyistä tilannetta, ja sen valmiuksia suorittaa projekti. Siinä pureudutaan paljon syvemmälle itse liiketoiminnan asettamiin kriteereihin ja tavoitteisiin ja tarkkaillaan miten käytettävissä olevat resurssit sopivat niihin. Raportointi perustuu faktapohjaiseen analyysiin yrityksen tilasta ja siitä tuotetaan seuraavat neljä raporttia:

Resurssien inventaario: Raportti käsittää kaiken projektin resursointiin liittyvän faktan. Siihen voidaan liittää tarvittavat henkilöt, koneet, data ja ohjelmistot. Resursseja ei listata vain tiedonlouhinnan osalta, vaan kattavasti kaikki projektiin saatavilla olevat resurssit.

Vaatimukset, olettamukset ja rajoitteet -raportti: Projektiin liittyvät vaatimukset, esimerkiksi jos projektin täytyy täyttää tiettyjä lain pykäliä, tai esimerkiksi tietoturvaan liittyvät vaatimukset. Jos projektissa esiintyy olettamuksia esimerkiksi jonkin lain tulkinnasta tai rajoitteita datan sensitiivisyyden osalta tulee nämä asiat kirjata tähän raporttiin. Tärkeimpänä on varmistaa pääsy dataan.

Terminologia: Raportti sisältää listauksen liiketoiminnan käyttämästä terminologiasta, jotka ovat relevantteja projektin osalta ja niiden selitykset.

Kustannukset ja hyödyt: Raportti kustannuksista ja projektin laskennallisista hyödyistä liiketoiminnan näkökulmasta. Arvio projektin kustannusten määrästä suhteutettuna projektin hyötyihin.

Tiedonlouhinnan tavoitteiden määrittelyssä tuotetaan kaksi raporttia, jotka kuvaavat tiedonlouhinnan tavoitteita ja onnistumiskriteerejä:

Tiedonlouhinnan tavoitteet: Määritellään tiedonlouhinnan tuotokset, mikä sisältää määrittelyn tuotetuista raporteista, malleista ja prosessoidusta datasta. Myös mahdolliset analyysien esitykset on hyvä määritellä.

Tiedonlouhinnan onnistumiskriteerit: Määritellään tekniset kriteerit tiedonlouhintatyölle, jotka ovat oleellista saavuttaa liiketoiminnan tavoitteiden saavuttamiseksi. Käytetään mahdollisuuksien mukaan tarkkoja lukuja esimerkiksi mallin tarkkuusprosentti tai ennustetarkkuuden parannusprosentti.

Viimeinen osio liiketoiminnan ymmärtämisestä on projektisuunnitelman laatiminen, perustuen aikaisemmista kohdista saatuihin tietoihin. Tässä tehtävässä on kaksi alatehtävää, joista muodostuu kaksi raporttia:

Projektisuunnitelma: Määritellään projektille tarkka projektisuunnitelma, jossa kuvataan tarkasti jokaisen vaiheen vaatima aika, vaiheen toteutukseen tarvittavat henkilöt (esimerkiksi erikoisosaajat aihealueesta), tarvittavat resurssit ja suunnitellut tuotokset. Lisäksi kuvataan tehtävien väliset riippuvuussuhteet ja luodaan tehtävien suorittamisjärjestys. Kuvataan kaikkien iteratiivisten osioiden sykli ja toistuvuuskerrat.

Alustava työkalujen ja tekniikoiden valinta: Raporttiin kuvataan tiedonlouhinnan tavoitteiden saavuttamiseksi vaadittavat työkalut ja tekniikat. Kuvataan omat resurssit ja nostetaan ylös ne tarpeet, mitä ei ole saatavilla tai mitkä ovat puutteellisia.

4.3.2 Datan ymmärtäminen

CRISP-DM:n toinen vaihe käsittää projektiin valitun datan ymmärtämisen. Siinä käydään läpi kaikki ne datat, jotka liiketoiminnan ymmärtämisessä on valittu projektin dataresursseiksi. Datan ymmärtämiseen lasketaan työvaiheena myös datan lataaminen työkaluun, jos datan ymmärtämiseen käytetään spesifiä työkalua. Jos dataa tulee myös useammasta lähteestä, niin vaiheessa tulee myös miettiä mahdollisten integraatioiden, eli datayhteyksien luomista ja niiden ajoittamista tiettyyn vaiheeseen. Datan ymmärtäminen koostuu neljästä eri tehtävästä, niiden alatehtävistä, sekä yhdestä yleisestä vaiheesta, jossa koostetaan alustava raportti datan keräämisestä.

Yleinen vaihe koostaa yhteen datan lähdetiedot ja niihin liittyvät asiat. Yleisestä vaiheesta tuotetaan yksi raportti, joka on kuvattu alla.

Alustava datan keruu: Raportissa kuvataan kaikki suunnitellut datalähteet, niiden sijainti, ja metodit, joilla data kerätään lähteistä. Raporttiin koostetaan myös tiedot kaikista ongelmatilanteista, joita datan keräämisessä kohdataan. Ongelmaraportoinnilla varaudutaan tulevaisuuteen. Jos projekti toistetaan myöhemmin uudestaan tai samoihin datalähteisiin kohdistetaan toisia projekteja, niiden mahdolliset ongelmat ja ratkaisut on jo valmiiksi dokumentoitu.

Datan ymmärtämisen ensimmäinen tehtävä on itse datan kuvaaminen. Datan kuvaamisella tarkoitetaan datan ylimpään kerrokseen kohdistunutta havainnointia, jolla koitetaan vastata kysymyksiin yleisistä datan ominaisuuksista. Datan kuvaamisesta tuotetaan yksi raportti.

Datan kuvausraportti: Kuvaa datan formaatin, datan määrän (esimerkiksi kuinka monta riviä ja saraketta datan jokaisessa taulussa on) sekä määrittää, onko kaikkien sarakkeiden nimet esimerkiksi löydetty. Raportissa arvioidaan myös ylätasolla, onko datan määrä esimerkiksi sellainen, että suunniteltu projekti voidaan aloittaa.

Toinen tehtävä datan ymmärtämisessä on datan tutkiminen. Datan tutkimisessa käsitellään tiedonlouhinnalle asetettuja kysymyksiä. Vaiheessa käytetään hyväksi erilaisia tekniikoita ja työkaluja, kuten datan visualisointia, jolla tarkoitetaan datasta löytyvien asioiden esittämistä visuaalisessa muodossa, esimerkiksi eri muuttujien, eli datan attribuuttien, summien piirtämistä pylväsdiagrammeina. Vaiheessa suoritetaan myös perinteisiä statistisia analyyseja, joissa käsitellään esimerkiksi datan keskihajontaa, keskiarvoja sekä vaihteluvälejä. Lisäksi saatetaan muodostaa aggregointeja, joiden tarkoituksena on esimerkiksi saada joillekin ajanjaksolle keskiarvoistettuja tuloksia, jotta ajanjaksoa voidaan ymmärtää paremmin suhteessa muihin ajanjaksoihin. Lisäksi statistisilla menetelmillä voidaan tutkia muuttujien välisiä korrelaatioita tai pienen muuttujajoukon muodostamien ryhmien korrelaatioita. Korrelaatiomenetelminä voidaan käyttää esimerkiksi Pearssonin korrelaatiota, otoskorrelaatiota tai ei-parametrisia korrelaatiokertoimia, kuten Kendalin korrelaatiokerroin tai Spearmanin järjestyskorrelaatiokerroin. Nämä analyysit voivat suoraan vastata tiedonlouhinnalle asetettuihin tavoitteisiin. Tulokset voivat myös

vaikuttaa datan kuvaukseen tai seuraavassa tehtävässä kuvattuun datan laaturaporttiin. Jos analyysit osoittavat huonoja tuloksia asetettuihin tavoitteisiin nähden, voidaan joutua palaamaan takaisin datan kuvaamisen raportointiin, tai jopa pohtia, tuleeko hakea uutta dataa tai transformoida eli muokata nykyistä dataa niin että, se vastaisi paremmin tiedonlouhinnan tavoitteisiin. Vaiheesta tuotetaan yksi raportti.

Datan tutkiminen: Kuvataan datan tutkimisessa tehtyjen analyysien tulokset, jotka sisältävät ensimmäiset havainnot, mahdolliset alustavat hypoteesit ja niiden vaikutukset jäljellä oleviin projektin osioihin. Raportin ymmärtämisen kannalta on tärkeää sisällyttää tehdyt datan visualisoinnit raporttiin, jotta raportin luettavuus ja ymmärrettävyys on parempi.

Kolmantena tehtävänä on datan laadun varmentaminen. Datan laadun varmentaminen on tärkeä prosessin vaihe. Väärin tehdyllä ja raportoidulla vaiheella voi olla dramaattisia seurauksia myöhemmissä vaiheissa, kun dataa aletaan mallintamaan. Datan laadun varmentamisessa tarkkaillaan datan kokonaisuutta, eli kattaako data kaikki mahdolliset vaaditut tapaukset. Laadun varmistuksessa tarkkaillaan myös virheitä Oleellisina kysymyksinä on, sisältääkö data virheitä vai onko se virheetöntä. Jos data sisältää virheitä, niin mitä nämä virheet ovat ja voidaanko ne korjata? Myös datan eheys varmennetaan datan laadun varmennuksen vaiheessa. Tässä vaiheessa suoritetaan analyysi mahdollisista puuttuvista arvoista datassa sekä siitä, missä attribuuteissa nämä puuttuvat arvot ovat, ja kuinka paljon niitä on. Tarkkaillaan myös sitä, miten puuttuvat arvot esiintyvät. Yleensä puuttuvilla arvoilla on kaksi tyypillistä esiintymismuotoa. Joko ne ovat datassa tyhjiä datapisteitä, eli matriisin paikkoja, tai sitten ne on merkattu puuttuvaksi arvoksi NaN- tai null-arvoilla. Tästä tehtävästä ei tuoteta varsinaista raporttia vaan saatuja havaintoja käytetään tehtävässä neljä.

Tehtävä neljä käsittää **Datan laadun raportoinnin**. Tehtävässä ei ole varsinaisia alatehtäviä vaan siinä käytetään hyödyksi aikaisemmista tehtävistä saatuja tuloksia. Raportissa listataan saadut tulokset datan laadun varmentamisesta ja sitä verrataan datan ymmärtämisen ja liiketoiminnan ymmärtämisen asettamiin tavoitteisiin. Jos näissä esiintyy ristiriitoja, niin niihin ehdotetaan raportissa ratkaisuja.

4.3.3 Datan valmistelu

CRISP-DM:n kolmas vaihe on datan valmistelu. Vaiheen tarkoituksena on valita projektiin sopiva data, valmistella se ennakkoon mallinnusta varten, rakentaa lisää dataa tarvittaessa ja yhdistää eri datat yhteen, jos se on tarpeellista. Datan valmistelun vaihe jakaantuu neljään tehtävään, josta jokaisella on omia alatehtäviään.

Ensimmäinen tehtävä on datan valinta. Tehtävässä tehdään päätökset siitä, mitä dataa projektissa tullaan lopulta käyttämään saatavilla olevista datoista. Datan valintaan voidaan asettaa kriteereitä, jotka muodostuvat aikaisemmista CRISP-DM:n vaiheista. Datan valinnan kriteereinä voivat toimia datan relevanttius asetettuun tavoitteeseen nähden, datan laatu sekä joukko erilaisia teknisiä ehtoja kuten datan määrän rajoitteet tai datan tyyppi. Huomattavaa on, että datan valinnalla tarkoitetaan sekä attribuuttien, eli sarakkeiden valintaa, että itse rivien valintaa. Tehtävästä koostetaan yksi raportti.

Valintojen perusteet: Raportissa kuvataan kaikki datan valinnat, niin otto kuin rajaukset sekä niiden perustelut. Raportti muodostaa siis säännöt datan valinnalle.

Seuraava tehtävä datan valmistelussa on datan puhdistaminen. Tehtävän tarkoituksena on nostaa valitun datan laatua ja parantaa sen käytettävyyttä itse projektissa. Datan puhdistamiseen käytetään useita menetelmiä datan laatuongelmista riippuen. Menetelminä voidaan käyttää esimerkiksi tietyn puhtaan, eli laadukkaan osajoukon eristäminen kokonaisdatasta. Muita menetelmiä on puuttuvien arvojen poistaminen, niiden keskiarvoistaminen, sopivien vakioiden lisääminen dataan tai mallinnuksen avulla simuloiminen. Mallinnuksen avulla simuloimisella tarkoitetaan puuttuvan datan luomista mallintamalla muusta datasta. Menetelminä mallinnuksessa käytetään tarkoitukseen sopivia statistisia menetelmiä, kuten logistista regressiota tai lineaarista regressiota. Tästä tehtävästä luodaan yksi raportti.

Datan laatu: Raportissa kuvataan kaikki menetelmät, joita käytettiin datan laatuongelmien poistamiseksi. Raportissa myös arvioidaan tehtyjen prosessien vaikutusta datan soveltuvuuteen itse projektin suhteen.

Kolmas tehtävä datan valmistelussa on tarvittavan datan rakentaminen. Tehtävässä johdetaan uusia muuttujia vanhoista, luodaan uusia rivejä dataa tai muunnetaan

aikaisempien muuttujia. Datan rakentaminen ei ole aina välttämätöntä, jos käytössä oleva data tarjoaa kaikki tarvittavat muuttujat ja tarvittavan variaation itse tapahtumista. Tehtävästä muodostetaan kaksi raporttia, jotka kuvaavat vaiheita, kuinka uudet datapisteet on luotu.

Johdetut muuttujat: Raportissa kuvataan kaikki johdetut muuttujat ja niiden lähdetiedot. Raportin tarkoitus on säilyttää tieto siitä, miten uusia muuttujia on muodostettu ja näin säilyttää datan läpinäkyvyys, joten lähdemuuttujien ja kaikkien välivaiheiden kuvaaminen on tärkeää.

Luodut rivit: Raporttiin kuvataan uusien rivien luonti samalla tavalla kuin johdetut muuttujat. Uusien rivien luomisessa ei yleensä käytetä lähteenä aikaisempia datan tietoja, vaan luodaan kokonaan uusia tapauksia. Jos näitä rivejä luodaan, raporttiin kuvataan myös perustelut siitä, miksi uusia rivejä täytyi luoda.

Datan valmistelun neljäs ja viimeinen tehtävä on datan integrointi. Datan integroinnilla tarkoitetaan projektissa käytettyjen datataulujen yhdistämistä, tai datan aggregoinnilla saatuja uusia tauluja. Datan aggregoinnilla voidaan luoda esimerkiksi päivätasolla mitattavan kaupan kuukausien keskimääräiset myynnit, ja näin luodaan kuukausikohtainen taulu päiväkohtaisen taulun sijasta. Tehtävästä muodostetaan yksi raportti.

Datan yhdistäminen: Raportissa kuvataan yhdistettyjen datojen lähdetiedot, miten yhdistäminen on tehty ja mikä sen lopputulos on. Myös aggregoiduista tauluista muodostetaan samanlainen kuvaus, jossa kerrotaan, minkä attribuutin/attribuuttien mukaan taulu on aggregoitu ja mitä tietoja uuden taulukon attribuuteiksi tulee.

4.3.4 Mallintaminen

Mallintaminen on CRISP-DM:n neljäs vaihe, ja sillä tarkoitetaan muokatun ja puhdistetun datan mallintamista malliksi, joka täyttää tiedonlouhinnan tavoitteet ja vastaa liiketoiminnan asettamiin tavoitteisiin. Mallinnuksella tarkoitetaan esimerkiksi koneoppivan, tai syväoppivan algoritmin opettamista datasta. Opetetulla mallilla voidaan ennustaa haluttua tähtäinmuuttujaa. Mallintaminen koostuu neljästä tehtävästä, joilla on useita alatehtäviä.

Mallintamisen ensimmäinen tehtävä on mallinnustekniikan valinta. Mallinnustekniikan valinnalla tarkoitetaan lopullista valintaa eri mallien välillä. Mallinnustekniikoita voidaan valita myös useita, mutta silloin mallinnuksen kaikki tehtävät on toistettava jokaiselle mallinnustekniikalle erikseen. Mallinnustekniikan valinnassa tuotetaan kaksi raporttia.

Mallinnustekniikka: Raporttiin kuvataan tarkasti käytettävä mallinnustekniikka ja mallinnuksen ominaisuudet.

Mallinnuksen olettamukset: Kuvataan kaikki olettamukset, joita mallinnustekniikka tekee. Jotkut mallinnustekniikat vaativat esimerkiksi, ettei datassa ole puuttuvia arvoja, tai sen käyttämä datan jakauma asettuu tietylle välille. Kaikki tällaiset olettamukset tulee kirjata ylös.

Toisena tehtävänä on mallinnuksen testauksen suunnittelu. Mallintestaussuunnitelmalla varmistetaan, että rakennettua mallia voidaan testata ja sillä saadut tulokset voidaan validoida. Testaussuunnitelmaan kuvataan testauksessa käytetyt metodit ja testauksen proseduuri. Proseduurilla tarkoitetaan sitä mallin testausmenetelmien ketjua, jonka avulla mallin tarkkuus voidaan varmentaa. Testausmenetelmillä tarkoitetaan kuhunkin malliin soveltuvia testausmenetelmiä, esimerkiksi klassifioivissa malleissa käytetään yleisesti erilaisia virherajoja. Mallin testaussuunnitelmaan kuuluu myös aiotun opetus- ja testidatan jaottelusuhde. Tehtävästä tehdään yksi raportti.

Testaussuunnitelma: Raporttiin kuvataan mallin koulutuksen, testauksen ja validoinnin suunnitelma. Primäärikomponenttina raportissa on koulutus-, testaus ja validointidatan suhteet sekä käytettävät mallin tarkkuuteen viittaavat menetelmät.

Kolmantena mallinnuksen tehtävänä on itse mallin rakentaminen valitulla/valituilla mallinnustekniikoilla, missä käytetään aikaisemmissa vaiheissa valmisteltua ja validoitua dataa. Mallin rakentamisesta muodostetaan yksi raportti.

Malliraportti: Raporttiin sisällytetään kuvaus käytetyistä malleista, jotka sisältävät kattavan läpikäynnin mallien ominaisuuksista ja rakennusvaiheessa sattuneista ongelmista sekä niiden ratkaisuista. Raporttiin kuvataan myös mallien käyttämät parametrit ja valittujen parametrien valintaperusteet. Malliraporttiin liitetään myös itse

mallit. Jos mallit on mahdollista saada tiedostona, niin nämä tiedostot liitetään osaksi raporttia.

Viimeisenä mallinnuksen tehtävänä on mallin arviointi. Mallin arvioinnissa tehdään johtopäätöksiä toimiala osaamisen, tiedonlouhinnan kriteerien ja testaussuunnitelman mukaan. Tavoitteena on arvioida mallinnussovellusta ja sen löydöksiä tekniseltä näkökulmalta. Ensimmäisen arvioinnin jälkeen arvioidaan tiedonlouhinnan tulosten soveltuvuutta liiketoiminnan kontekstiin. Mallin arvioimisessa keskitytään vain mallin itsensä arviointiin, kun taas seuraavassa CRISP-DM-mallin vaiheessa keskitytään koko projektin arviointiin. Mallin arvioinnista tuotetaan yksi raportti.

Mallin arviointi: Listataan kaikkien mallien ominaisuudet, kuten niiden tarkkuus, koulutusaika ja virheet. Verrataan niitä toisien mallien vastaaviin ominaisuuksiin, ja ne järjestetään toisiinsa nähden paremmuus järjestyksessä valittujen kriteerien mukaan. Mallinnuskierrosten jatkuessa, myös mallien uudistetut parametrit kirjataan mallin arviointiraporttiin, jotta nähdään, millä asetuksilla malli on kehittymässä mihinkin suuntaan.

4.3.5 Arviointi

CRISP-DM:n arviointi vaiheessa arvioidaan koko projektin onnistuneisuutta sille asetettuihin liiketoiminnan ja tiedonlouhinnan tavoitteita vasten. Vaiheen aikana arvioidaan myös itse prosessi ja määritetään arvioinnin perusteella projektin seuraavat askeleet. Vaihe koostuu kolmesta tehtävästä ja jokaisella tehtävällä on omia alatehtäviä.

Ensimmäinen tehtävä on saatujen tulosten arviointi. Tulosten arvioinnin avainkysymyksenä on selvittää, onko olemassa jotain liiketoiminnasta tulevaa syytä, miksi malli ei olisi soveltuva täyttämään liiketoiminnan tarpeita. Toinen arviointimenetelmä on siirtää malli suoraan tuotantoon ja käyttää sitä tuotannossa osana järjestelmää ja arvioida sen tuottamia tuloksia. Tämän arviointimenetelmän huonona puolena kuitenkin on sen kalleus, sekä vaadittu aika. Arvioinnissa myös tarkastellaan kaikkia niitä vastauksia, mitä sekundäärisiin tutkimuskysymyksiin on tiedonlouhintaprojektin aikana pystytty saamaan vastaus, ja onko tullut esille uusia sekundäärisiä kysymyksiä tai ratkaisuja mallinnusprosessin aikana, jotka eivät

kuitenkaan liity varsinaiseen primääriseen liiketoiminnan asettamaan ongelmaan. Arviointivaiheesta tuotetaan yksi raportti.

Tulosten arviointi ja hyväksytyt mallit: Raportissa vedetään yhteen kaikki tiedonlouhinnalla saavutetut tulokset ja niitä verrataan liiketoiminnan ja tiedonlouhinnan tavoitteisiin. Raportti sisältää myös lopullisen lauselman, joka käsittää arvion kaikista projektin vaiheista riippumatta siitä, täyttikö projekti liiketoiminnan asettamia tavoitteita vai ei. Raporttiin lisätään myös listaus kaikista niistä malleista, jotka täyttävät liiketoiminnan asettamat tavoitteet. Näistä malleista muodostuu hyväksyttyjen mallien joukko.

Toisena tehtävänä on prosessin arviointi. Prosessilla tarkoitetaan tässä yhteydessä koko tiedonlouhintaprosessia. Tarkoituksena on selvittää, onko prosessin aikana löytynyt sellaisia vaiheita tai asioita, joiden tärkeyttä ei osattu määritellä projektin suunnitteluvaiheessa kunnolla. Prosessin arvioinnin aikana selvitetään myös, onko mallinnuksen ja projektissa tehtyjen löydösten laatutaso sellainen, että mallin tuloksiin voi luottaa ja ettei mitkään muutkaan projektin vaiheet ole päässeet sotkemaan mallinnuksen luotettavuutta. Lisäksi arvioidaan myös projektin laillisuusperiaatteita ja hyvän tavan mukaisia käytänteitä, esimerkiksi sitä, käytettiinkö mallinnuksessa vain sellaista dataa, jota sai käyttää mallinnukseen ja prosessointiin. Prosessin arvioinnista koostetaan yksi raportti.

Prosessin arviointi: Raporttiin kuvataan kaikki prosessin vaiheet ja arvioidaan niitä kriittisesti, korostetaan niitä aktiviteettejä, jotka näyttelivät suurempaa roolia kuin oli ennakkoon suunniteltu tai joihin ei pystytty ennakkoon varautumaan. Lisäksi kirjataan ylös kaikki ne prosessin vaiheet, jotka jouduttiin toistamaan.

Kolmantena tehtävänä arviointivaiheessa on määrittää projektin seuraavat askeleet. Tehtävässä päätetään projektista saatujen tietojen ja asetettujen tavoitteiden mukaan se, onko tiedonlouhinnalla saavutettu riittävä vastauskyky liiketoiminnan tarpeeseen, ja jos on, niin voidaanko jatkaa tuotteistamiseen. Vaiheessa voidaan myös saatujen tietojen perusteella päättää, ettei projekti ole saavuttanut sille asetettuja tavoitteita ja sen tulee käynnistää uusi iteraatiokierros, tai projekti voidaan myös kokonaan lopettaa, jos nähdään, ettei asetettuihin tavoitteisiin tai ongelmaan voida saada kannattavalla tavalla ratkaisua. Vaiheesta tuotetaan yksi raportti.

Projektin jatkon määrittely: Raporttiin listataan kaikki mahdolliset vaihtoehdot projektin jatkolle, ja niiden mahdolliset vaikutukset sekä perustelut niin päätösvaihtoehdon puolesta kuin sitä vastaan. Päätösvaihtoehtojen arvioinnin jälkeen tehdään itse päätös, joka myös kirjataan raporttiin.

4.3.6 Tuotteistaminen

Tuotteistamisella tarkoitetaan kehitetyn mallin viemistä osaksi yrityksen liiketoimintaa, ja sen jatkuvan kehityksen, monitoroinnin sekä huollon suunnittelua. Tuotteistamisvaiheessa tehdään myös lopullinen raportti kaikista projektin vaiheista, jonka jälkeen koko projektia vielä arvioidaan. Tuotteistaminen ei tapahdu projektissa kuin kerran, ja mahdollisen iteraation kohdalla CRISP-DM:stä tehdään kaikki muut kohdat paitsi tuotteistaminen. Silti tuotteistaminen on äärimmäisen tärkeä projektin osa, ja ilman sitä ei aikaisemmalla työllä ole hyötyä liiketoiminnalle. Vaihe jakaantuu neljään tehtävään, joista jokaisella on omia alatehtäviä.

Ensimmäinen tehtävä on tuotteistamisen suunnittelu. Tuotteistamisen suunnittelussa arvioidaan arviointivaiheen tuloksia ja niiden pohjalta laaditaan suunnitelma tehtävästä tuotteistamisesta. Suunnitelmassa arvioidaan eri tapoja tuotteistaa saadut mallit ja muodostetaan parhaista vaihtoehdoista strategia. Tehtävästä muodostetaan yksi raportti.

Tuotteistamisen suunnitelma: Raporttiin kuvataan tuotteistamisen strategia, josta selviää tarkasti suunnitelma siitä, miten tehdyt mallit istutetaan yrityksen liiketoimintaan. Jos liiketoiminnan tavoitteena oli esimerkiksi saada uusi näkymä yrityksen ensi kuun ennusteista graafisesti, tulee tuotteistamisen suunnitelmaan laatia kuvaukset esimerkiksi vaadittavasta sovellusarkkitehtuurista sekä miten itse mallit tullaan upottamaan tähän sovellusarkkitehtuuriin.

Toisena tehtävänä tuotteistamisessa on suunnitella mallin monitorointi ja huolto. Mallien siirtyessä tuotteissa yrityksen jokapäiväiseen toimintaan, on niiden toiminnan jatkuvuus äärimäisen tärkeää yrityksen liiketoiminnan kannalta. Huolellisesti laadittu huolto- ja monitorointisuunnitelma ehkäisee vikatilojen pitkittymistä sekä auttaa havaitsemaan mahdollisen mallin väärän toiminnan ajoissa, jolloin väärin toimivan mallin antamien tulosten vaikutus yrityksen liiketoimintaan voidaan minimoida. Monitorointi- ja

huoltosuunnittelussa määritellään vastuuhenkilöt, heidän toimenkuvansa, ja raja-arvot, jolloin malli ei toimi enää niin kuin sen pitäisi. Lisäksi siinä määritellään esimerkiksi huoltoajat, jolloin mahdollisilla huoltotoimenpiteillä on mahdollisimman vähäinen vaikutus yrityksen liitetoimintaan. Monitoroinnin ja huollon suunnittelusta tuotetaan yksi raportti.

Monitoroinnin ja huollon suunnitelma: Raporttiin kuvataan tarkasti edellä mainitut huollon ja monitoroinnin osa-alueet niin, että ne on kaikille osapuolille selvät. Suunnitelmaan lisätään myös kaikki tieto siitä, miten näitä suunniteltuja toimenpiteitä käytännössä toteutetaan.

Tuotteistamisen kolmantena tehtävänä on projektin loppuraportin luominen. Loppuraportti voi käsitellä vain projektin yhteenvetoa ja siitä saatuja kokemuksia, tai sitten se voi olla kokonaisvaltainen kokonaisuus kaikista tiedonlouhinnan vaiheista ja tuloksista. Tehtävästä tuotetaan kaksi raporttia.

Loppuraportti: Raporttiin kuvataan valitulla tavalla kaikki projektin löydökset. Se sisältää kuvauksen siitä, miten nyt tehty projekti täydentää mahdollisesti aikaisemmin tuotettuja ratkaisuja, mitä projektilla saavutettiin ja millä tuloksilla.

Loppuesitys: Loppuesityksen tarkoituksena on tiivistää loppuraportin pääkohdat niin, että se voidaan tiiviisti ja selkeästi esittää asiakkaalle.

Tuotteistamisen neljäs tehtävä on projektin arviointi. Siinä tutkaillaan projektin johtamisen näkökulmasta kaikkia aspekteja siitä, miten projektissa on onnistuttu ja missä epäonnistuttu. Arvioinnissa otetaan kantaa myös omien prosessien parantamiseen. Jos joku prosessin osa vaatii parannusta, siitä tehdään kehitysehdotus, jossa kuvataan parannettavan prosessin osan heikkouksia, sekä toimenpiteitä, joilla prosessia saadaan paremmaksi. Raporttiin liitetään tiedoksi myös kaikki projektissa havaitut sudenkuopat ja yllätykset, jotta vastaavissa projekteissa niitä osattaisiin välttää paremmin. Tehtävästä kirjoitetaan yksi raportti.

Projektin arviointi: Kuvataan projektin kaikki vaiheet täydellisesti ja pyritään huomioimaan kaikki muutosta vaativat tekijät sekä tekijät, jotka onnistuivat projektissa

hyvin. Raportin ideana on koota kattavasti kaikki projektin havainnot yhteen raporttiin, jotta jatkossa voidaan helpommin varautua vastaaviin projekteihin.

4.4 Chatbot

Chatbotilla tarkoitetaan ohjelmistoa, jolla tekoälyn avulla simuloidaan ihmisen luontaisella kielellä käytyä keskustelua chat-applikaation välityksellä. Chatbotin tarkoitus on toimia käyttöliittymänä ihmisen ja ohjelmiston välillä, ja sillä voidaan korvata osaksi esimerkiksi asiakaspalvelijan tekemää työtä. Chatbotit käyttävät usein hyödykseen erilaisia koneoppimisella tehtyjä malleja, kuten luontaisen kielen prosessointia, joiden avulla esimerkiksi tekstin konteksti ja semantiikka voidaan opettaa paremmin botille. Chatbotit koostuvat pääasiassa kolmesta eri osiosta: intentistä, entiteetistä sekä dialogista. [14.]

Intentillä tarkoitetaan esimerkinomaista lausetta, jolla opetetaan chatbotille konteksti, mistä asiayhteydestä puhutaan. Entiteetillä taas tarkoitetaan spesifistä avainsanaa, joka linkkaa kontekstin tiettyyn tekemisen kohteeseen. Dialogissa taas muodostetaan intenteistä ja entiteeteistä itse keskustelun logiikka. Logiikalla tarkoitetaan sitä, miten intentit ja entiteetit ohjaavat käyttäjän vastausten perusteella keskustelua eteenpäin ja antavat kuhunkin kysymykseen vastaukseen ennalta sovitun mukaisesti. [14.]

Chatbot käyttää hyväkseen tekstianalytiikan sovellutuksia, joilla se analysoi käyttäjän sille syöttämän tekstin. Näitä tekstianalytiikan sovellutuksia ovat muun muassa luontaisen kielen prosessointi (Natural language processing) sekä luontaisen kielen klassifiointi (Natural Language Classification.) Se tunnistaa tekstistä sille opetetut intentit ja entiteetit, ja muodostaa niiden pohjalta vastauksen. [14.] Perinteisen chatbotin kyvykkyydet ovat riittäviä ohjaavaan toimintaan, esimerkiksi nettisivun aputoimintana, mutta ne eivät täytä käyttäjien asettamia vaatimuksia, jos chatbotilla halutaan esimerkiksi korvata yrityksen asiakaspalvelijoita. Perinteinen chatbot toimii hyvin informaation välitykseen, ja sillä voidaan simuloida käytäviä keskusteluja erinomaisesti, missä tarkoituksena on käytännönläheisesti opettaa kysymään tiettyjä kysymyksiä. [14.]

Kehittyneempään chatbottiin liitetäänkin kolmansia sovellutuksia täydentämään chatbotin kyvykkyyksiä. Näiden ominaisuuksien avulla chatbot voi muun muassa antaa

käyttäjälle koneoppivaan predikatiivisuuteen pohjautuvia vastauksia, tai se voi esimerkiksi hakea käyttäjälle vastauksen tuotteen käyttöohjekirjasta. [15.]

4.5 Git-versionhallintajärjestelmä

Git on käytetyin moderni versionhallintajärjestelmä. Sen on kehittänyt Linus Torvalds vuonna 2005. Git perustuu aktiivisesti ylläpidettyyn Open Source -projektiin. Monet ohjelmistojenkehitysprojektit sekä kaupalliset että ei-kaupalliset, nojaavat versionhallinnassaan Gittiin. Git perustuu hajautettuun arkkitehtuuriin. Hajautetun arkkitehtuurin ansiosta Git on esimerkki DVCS:stä (Distributed version control system). Git ei sisällä vain yhtä tallennuspaikkaa koko muutoshistorialle, vaan jokainen käyttäjä käyttää täydellistä kopiota muutoshistoriasta. Gitillä on monia vahvuuksia verrattuna muihin ei-hajautettuihin versiohallinnanjärjestelmiin. Gitin vahvuuksia ovat tehokkuus, turvallisuus sekä joustavuus. Nämä ominaisuudet ovat olleet peruslähtökohtia Gitin suunnittelussa ja kehityksessä. [16.]

Gitin perustoiminnot: julkaisu, haaroitus, yhdistäminen sekä vertailu on suunniteltu tehokkuuden näkökulmasta. Git ei huomioi versiohistoriassa tiedoston nimeä, vaan vertailee sen sisältöä aikaisempiin versioihin. [16.]

Gitin turvallisuus perustuu käytössä olevaan SHA1-kryptausmenetelmään, jonka avulla kaikki Gitissä oleva tieto on kryptattu. Kryptauksen alaisuuteen kuuluu niin itse tietosisältö, versiomerkinnät, tiedostojen väliset suhteet, julkaisut sekä objektit. [16.]

Gitin joustavuudella tarkoitetaan yleisesti ei-lineaaristen kehittäjien työnkulkujen tukemista, skaalautumista sekä isoihin että pieniin projekteihin, ja sitä, että useat muut kehitystyökalut tukevat Gittiä. [16.]

Gittiä käytettäessä jokainen kehittäjä voi julkaista oman kehitystyönsä projektiin, josta se jaetaan toisten käyttäjien kopioihin ja päivitetään tehdyt muutokset kullekin käyttäjälle. Näin jokainen käyttäjä saa tehdyt muutokset suoraan käyttöönsä.

4.6 Oppimisalusta

Oppimisalustalla tarkoitetaan web-pohjaista ratkaisukokonaisuutta, joka yhdistää opetuksen tarjoajan, oppijat, sekä kolmannen osapuolen tekijät, esimerkiksi ulkoisen materiaalin tarjoajat. Oppimisalusta on usein verkko-oppimisympäristö ja määritelmän mukaan alustan tulee tarjota interaktiivinen online-palvelu, joka sisältää informaatiota, työkaluja, sekä materiaalia, joka tukee ja vahvistaa koulutuksellista tarjontaa. Sen tulee olla kokonaisvaltainen sekä kattava ratkaisu, joka mahdollistaa turvallisen, web-pohjaisen opiskelun intuitiivisella käyttöliittymällä. [17.]

Oppimisalustan sisältö voidaan karkeasti jakaa neljään opiskelua tukevaan sisältökokonaisuuteen.

Sisällönhallinta: Sisällönhallinnalla pidetään alustan sisältämä oppimismateriaali järjestyksessä. Sisällönhallinnan avulla tuodaan koulutukseen osallistuvalle tarpeellinen elektroninen oppimismateriaali saataville. Sisällönhallinta voi myös tukea sisällön tuottamista. Tämän lisätoiminnon avulla koulutuksen suunnittelijat voivat tuoda sisältöä nopeasti saataville tai päivittää vanhaa koulutussisältöä. [18.]

Opetussuunnitelman suunnittelu ja esittäminen: Opetussuunnitelman suunnittelun ja esittämisen työkaluilla tuodaan koulutuksen sisältö rakenteellisena esille koulutettavalle. [18.]

Kommunikaatio: Kommunikaatiotyövälineillä tarkoitetaan erilaisia viestimenetelmiä, joiden avulla tuetaan koulutettavan opiskelua ja tarjotaan koulutettavalle mahdollisuus lähettää kysymyksiä niin koulutuksen vetäjälle kuin muillekin koulutukseen osallistuville henkilöille. Kommunikaation välineinä voi toimia esimerkiksi keskustelupalsta, sähköposti tai muut suoran viestinnän tai pikaviestinnän sovellutukset. [18.]

Hallinto: Hallinnon työkaluilla tarkoitetaan niitä työkaluja, joilla pidetään kirjaa koulutuksen tarjoajasta, koulutuksen vetäjistä, koulutettavista sekä heidän edistymisestään, sekä arvosanoistaan, jos koulutuksessa käytetään arviointia. Tämän lisäksi hallinnontyökaluihin voidaan laskea kuuluvan myös kaikki itse alustan hallintaan tarvittavat työkalut ja dokumentointi. [18.]

Interaktiiviset oppimisalustat on lähtökohtaisesti suunniteltu koulutusinstituutioiden tarpeisiin, mutta niitä voidaan tehokkaasti hyödyntää myös yrityksissä. Oppimisalusta mahdollistaa nopeasti päivittyvän ja helposti ylläpidettävän koulutustarjonnan, joka on skaalautuva erikokoisille koulutusryhmille. Yritysmaailmassa interaktiiviset koulutusalustat myös mahdollistavat joustavan kouluttautumisen työn ohessa itseohjautuvasti. Näin työntekijän on helpompi aikatauluttaa omaa työtään ja koulutustaan, mikä tarjoaa vapauden suorittaa koulutusta silloin, kuin se sopii hänen omaan aikatauluunsa. Tehokas oppimisympäristö on integroitu muihin yrityksessä käytettäviin prosesseihin, esimerkiksi projekteissa käytettyihin projektin prosessimalleihin.

Interaktiivinen oppimisalusta myös mahdollistaa personoidun koulutuksen, jossa omaan koulutustarpeeseen vastaavia oppimiskokonaisuuksia voidaan valita eri moduuleista, ja näin rakentaa oma tarpeenmukainen koulutuskokonaisuus.

4.7 Teoreettinen viitekehys

Teoreettisessa osiossa käsitellyt teemat tukevat tutkimuksen tavoitteita kokonaisvaltaisesti. Tavoitteen mukaisen interaktiivisen oppimisalustan luomiseksi on ymmärrettävä mistä moderni interaktiivinen oppimisalusta koostuu, mitä tietoa oppimisalusta tarjoaa tietosisällöllisesti sekä mitä uusia toiminnollisuuksia oppimisalustaan voidaan tuoda. Yrityksen tavoite on luoda datatieteilijöille sopiva oppimisalusta, ja tämän tieteenalan asettamat vaatimukset esitellään datatiede- sekä tiedonlouhintaosioissa. Teoriaosiossa käsitellyt oppimisalustan perusperiaatteet luovat vaatimuslistan oppimisalustan tarpeista, CRISP-DM malli luo tietosisällön pohjan sekä chatbot täydentää oppimisalustaa uudella näkökulmalla tehokkaaseen, interaktiiviseen oppimiseen.

5 Ehdotuksen rakentaminen asiakasyritykselle

5.1 Katsaus ehdotuksen rakenteesta ja validoinnista

Asiakkaalle rakennetussa sovelluksessa toteutetaan oppimisalusta, joka nojaa rakenteellisesti teoriaosiossa mainittuun CRISP-DM-prosessimalli. Oppimisalusta on verkkopohjainen ratkaisu, jossa yhdistyy teoria, käytännön harjoitukset sekä chatbotin avulla toteutettu interaktiivinen, asiakkaan kanssa käytävää keskustelua mallintava kokonaisuus. Ehdotuksen rakenteessa käytetään syksyllä 2018 Metropolia Ammattikorkeakoulussa järjestetyn liiketalouden analytiikkaprojektin kurssisisältöä, muistiinpanoja ja kurssilta saatua palautetta.

Ehdotus koostuu kahdesta osiosta. Ensimmäinen osuus käsittelee teoreettista näkökulmaa interaktiivisen oppimisalustan rakenteesta, ja toinen osa esittelee alustan työvälineet ja niiden käytön.

Valmis ehdotus validoidaan haastattelujen avulla asiakasyrityksen sisällä. Rakennettua oppimisalustaa koekäytetään sekä jo datatiedettä ammatikseen tekevien toimihenkilöiden toimesta sekä niiden, jotka ovat kiinnostuneita datatieteestä, mutta eivät sitä vielä oman toimenkuvansa puitteissa tee. Koekäytön perusteella suoritetaan haastattelut, joissa kerätään palautetta alustan toimivuudesta ja sen soveltuvuudesta käyttötarkoitukseensa.

5.2 Kurssi

Osana tutkimusta järjestettiin liiketoiminnan analytiikan kurssi Metropolia Ammattikorkeakoulussa syksyllä 2018. Kurssin tarkoituksena oli pilotoida teoriaosiossa esitellyn CRISP-DM-prosessimallin sopimista oppimisalustan tietorakenteen pohjaksi sekä oppimisalustassa käsiteltävien aiheiden järjestykseksi. Kurssilla saatujen kokemusten perusteella CRISP-DM-malli voidaan muokata koulutuksen viitekehykseen sopivaksi kokonaisuudeksi, jossa pystytään keskittymään opiskelijoiden tärkeimpinä sekä vaativimpina pitämiin asioihin.

Kurssi järjestettiin syksyllä 2018, ja se kesti yhteensä 14 viikkoa. Kurssi koostui yhdestä kolmen tunnin luennosta viikossa sekä opiskelijoiden omatoimisista harjoituksista, jotka liittyivät kyseisen viikon luennon aihekokonaisuuksiin. Omatoimiset harjoitukset jakautuivat ennen luentoa olevaan valmistelevaan tehtävään sekä konkreettiseen harjoitukseen luennon jälkeen.

Kurssin aihekokonaisuudet käsittelivät CRISP-DM-mallin aiheita, ja luennot oli jaettu otsikkotasolla kyseisten aiheiden mukaan. Luentojen aiheet käsiteltiin CRISP-DM:n mukaisessa järjestyksessä. CRISP-DM-mallin rakennetta täydennettiin yleisellä materiaalilla liittyen datatieteisiin sekä analytiikkaan, jotta kurssin sisältö ja tietorakenne pystyttiin paremmin jäsentämään ja kohdistamaan tuotantotalousinsinöörien opintoihin.

Kurssin aihealueet olivat seuraavat:

- analytiikka ja Datatiede tänään ja tulevaisuudessa
- business Intelligence ja analytiikka
- analytiikka ja CRISP-DM
- ennustava- ja Ohjeistava analytiikka
- koneoppiminen
- syväoppiminen
- neuroverkot
- tuotteistaminen.

Kurssille osallistuivat kolmannen vuoden kansainvälisen liiketoiminnan tuotantotalousinsinööriopiskelijat, ja kurssi oli osa heidän pääaineopintojaan. Kurssin kohdeyleisö valikoitui heidän opintojensa valmiusasteen takia. Kurssille osallistuneet opiskelijat olivat opintojensa loppuvaiheessa, ja näin ollen työelämässä toteutettavan koulutuksen sisällön pilotointi oli luonnollista heillä. Tällä saatiin hyvää palautetta itse tutkimuksen tavoitteisiin peilaten.

Kurssin rakenne onnistui hyvin, ja pilottikokeilua voidaan pitää onnistuneena. Kurssilta kerätyn palautteen mukaan kurssi oli laajuudessaan sopiva, ja uudet aiheet täydensivät aikaisemmin opittua. Kritiikki kohdistui joidenkin luentojen laajaan tietomäärään sekä kurssilla suoritettujen tehtävien haastavuuden tasalaatuisuuteen. Kurssin osallistujat olisivat toivoneet haastavampia harjoituksia.

Palaute tukee käsitystä pilottikokeilussa käytetyn rakenteen toimivuudesta vaihtoehtona interaktiivisen oppimisalustan tietosisällön rakenteen pohjaksi. Oppimisalustaa rakentaessa on kuitenkin otettava huomioon kurssipalautteessa ilmennyt kritiikki tehtävien vaikeustason vaihtelevuudesta.

5.3 Interaktiivinen oppimisalusta

Interaktiivinen oppimisalustan tarkoituksena on olla aktiivinen oppimisympäristö datatieteen opiskeluun projektiluontoisessa työympäristössä. Oppimisalustan rakentamisen perustana on käytetty teoriaosuudessa esiteltyjä aiheita, sekä kurssilla pilotoituja menetelmiä ja aihekokonaisuuksia.

5.3.1 Alustan osa-alueet

Alustassa käytetään luvussa 4.5 Oppimisalusta mainittua oppimisalustan rakennetta. Se koostuu tietosisällöstä, tehtäväosiosta sekä kommunikaatiovälineistä.

5.3.2 Tietosisältö

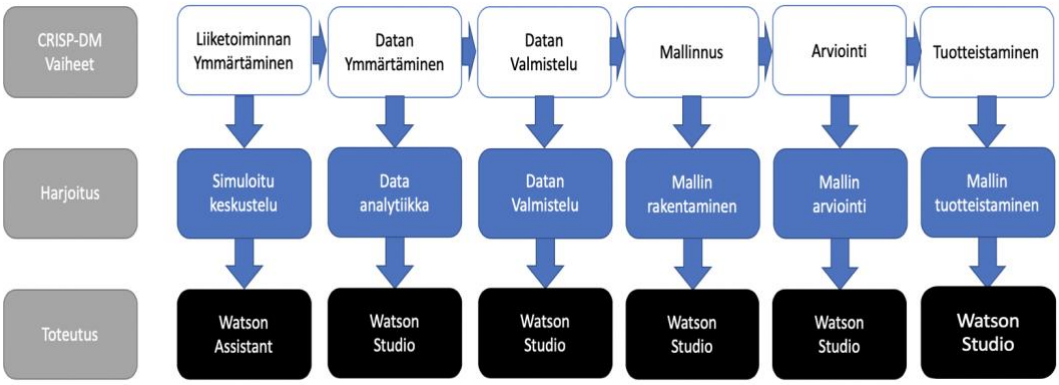
Oppimisalustan tietosisältö on jaettu luvussa 4.3 Cross-Industry Standard Process for Data-Mining (CRISP-DM) esitellyn CRISP-DM-mallin mukaisiin aihekokonaisuuksiin. Oppimisalustan vaatimuksena oleva sisällönhallinta on toteutettu staattisella verkkosivulla, jossa kukin mallin osa-alue on eritelty omaan osakokonaisuuteen. Osa-alueiden jakajana toimivat CRISP-DM-mallin pääotsikot, liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan preparointi, mallintaminen sekä arviointi ja tuotteistaminen. Teoriasisältöä täydennetään oppimisalustassa lisäksi ulkoisten artikkeleiden avulla, joilla tuetaan käytännönläheisesti itse teorian sisältöä. Käytännönläheisen tiedon ja esimerkkitapausten avulla teorian tukeminen koettiin analytiikkakurssilla hyväksi tavaksi sisäistää tietoa nopeammin ja selkeämmin.

5.3.3 Tehtävät

Oppimisalustan tehtäväosion rakenne on jaoteltu samoin kuten tietosisällön rakenne. Tehtäväsivut etenevät järjestyksessä kunkin vaiheen läpi, mikä tarjoaa katsauksen

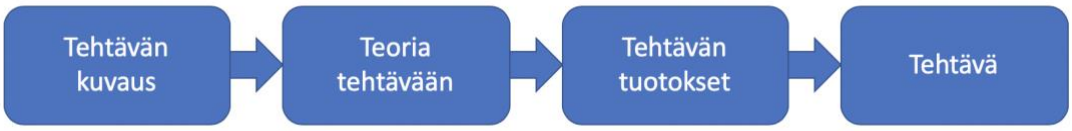
CRISP-DM-prosessimallin teoriaosuuteen, jonka jälkeen kunkin osion kohdalla on siihen suunniteltu harjoitustehtävä, jotka kokonaisuutena muodostavat datatiedeprojektin. Nykytila-analyysin perusteella toimihenkilöt kaipaavat enemmän tuotteiden ja työkalujen yhdistämistä keskenään. Kyseiseen asiaan on oppimisalustalla pyritty vastaamaan sillä, että annetut tehtävät, ja niiden ohjeet on tehty teknologiapohjaisesti. Ohjeet sisältävät samalla ohjeita itse työkalun käyttöön sekä siihen, miten datatiedettä tehdään käytössä. Näin oppimisalustalla on saatu yhdistettyä itse teoriaosaaminen sekä työkalujen osaaminen.

Kuvan 3 kaavio kuvaa CRISP-DM-mallin käyttöä tehtävärakenteen pohjana sekä harjoituksen toteutukseen käytettyä teknologiaa.



Kuva 3. CRISP-DM-mallin mukainen tehtävien jako, ja niiden toteutus

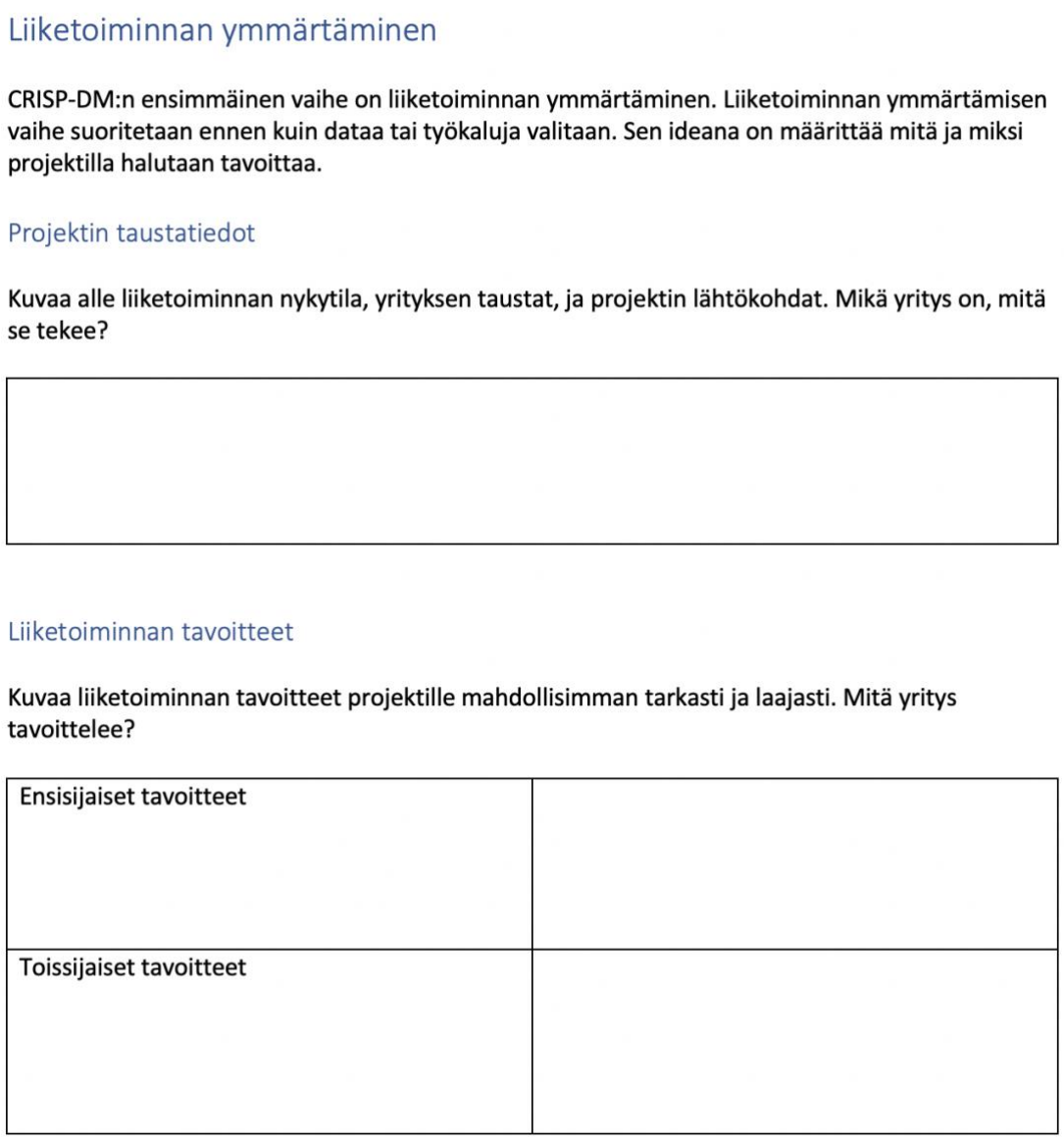
Kunkin CRISP-DM:n mukaisen osa-alueen tehtäväosio rakentuu oppimisalustassa kuvan 4 kaavion mukaisesti.



Kuva 4. Alustan tehtäväsivun rakennekaavio

Tehtäväsivuilla tehtävät ovat jaettu kahteen eri osa-alueeseen: CRISP-DM:n mukaiseen raportointiin sekä itse analytiikan harjoituksiin. Raportit vastaavat CRISP-DM-mallin vaatimuksia projektin aikana selvitettävistä asioista. Raportit on jaettu alustassa pienempiin palasiin kunkin osion osalle. Raporteilla pyritään vastaamaan

projektimuotoisen raportoinnin ja projektityön osaamisen tarpeeseen. Kuvassa 5 on esitelty Liiketoiminnan ymmärtäminen -raportin ensimmäinen sivu.



Liiketoiminnan ymmärtäminen

CRISP-DM:n ensimmäinen vaihe on liiketoiminnan ymmärtäminen. Liiketoiminnan ymmärtämisen vaihe suoritetaan ennen kuin dataa tai työkaluja valitaan. Sen ideana on määrittää mitä ja miksi projektilla halutaan tavoittaa.

Projektin taustatiedot

Kuvaa alle liiketoiminnan nykytila, yrityksen taustat, ja projektin lähtökohdat. Mikä yritys on, mitä se tekee?

Liiketoiminnan tavoitteet

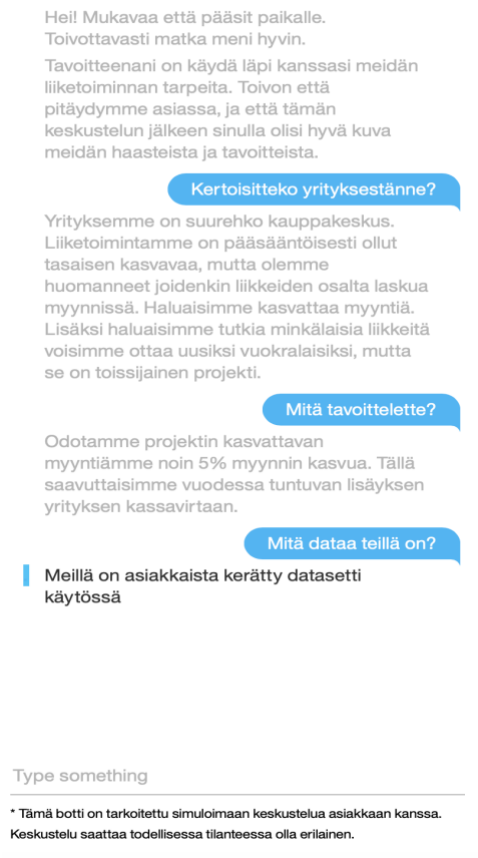
Kuvaa liiketoiminnan tavoitteet projektille mahdollisimman tarkasti ja laajasti. Mitä yritys tavoittelee?

Ensisijaiset tavoitteet	
Toissijaiset tavoitteet	

Kuva 5. Liiketoiminnan ymmärtämisen -raportin ensimmäinen sivu

Liiketoiminnan ymmärtäminen -osion yhteydessä käytetään myös luvussa 4.4 Chatbot esiteltyä chatbot-teknologiaa, joka on toteutettu IBM:n Watson Assistant -palvelulla. Chatbotin tarkoituksena on simuloida asiakkaan kanssa käytävää keskustelua, ja näin luoda autenttisempaa sisältöä koulutukseen. Simuloidussa keskustelussa käydään asiakkaan kanssa läpi heidän tarvettaan, vaatimuksiaan sekä rajoitteitaan, jotka

projektiin kohdistuu. Simuloitu keskustelu on turvallinen tapa käydä tekemisen kautta läpi käytännössä niitä kysymyksiä, joita todellisessa tilanteessa asiakkaalta täytyy kysyä.

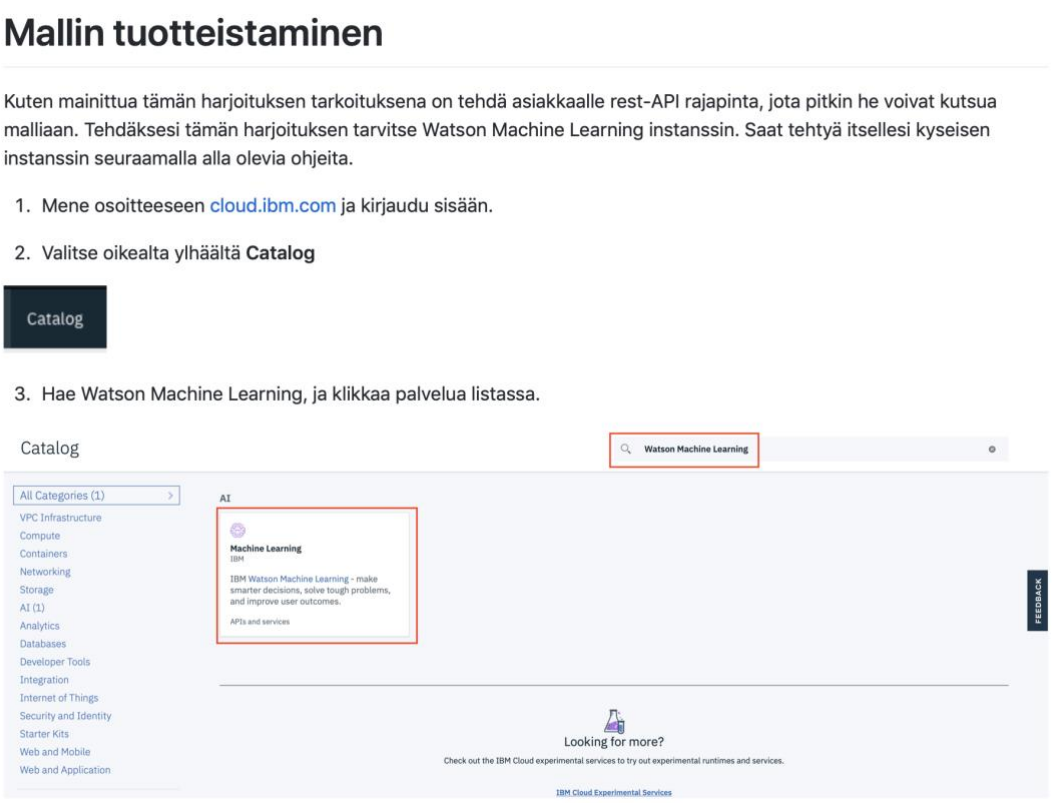


Kuva 6. Chatbot-keskustelu asiakkaan kanssa

Keskustelun päätteeksi kootaan raportti, joka ohjaa CRISP-DM-mallin mukaiseen dokumentointiin.

Itse analytiikan harjoitukset tehdään Watson Studio -työkalussa. Harjoitukset ovat itseohjautuvia, mutta niiden tekemiseen on tarjolla ohjeet. Itseohjautuva harjoittelu mahdollistaa sen, että oppimisalustaa pystyy käyttämään myös suoraan työelämässä projektin tukirakenteena. Se on myös adaptiivinen kunkin oppijan omaan taitotasoon

nähden. Perustason tehtävät, jotka alustaan on liitetty, ovat vaihe vaiheelta opastettuja. Kuvassa 7 käy ilmi harjoitusten ohjeiden tyyli.

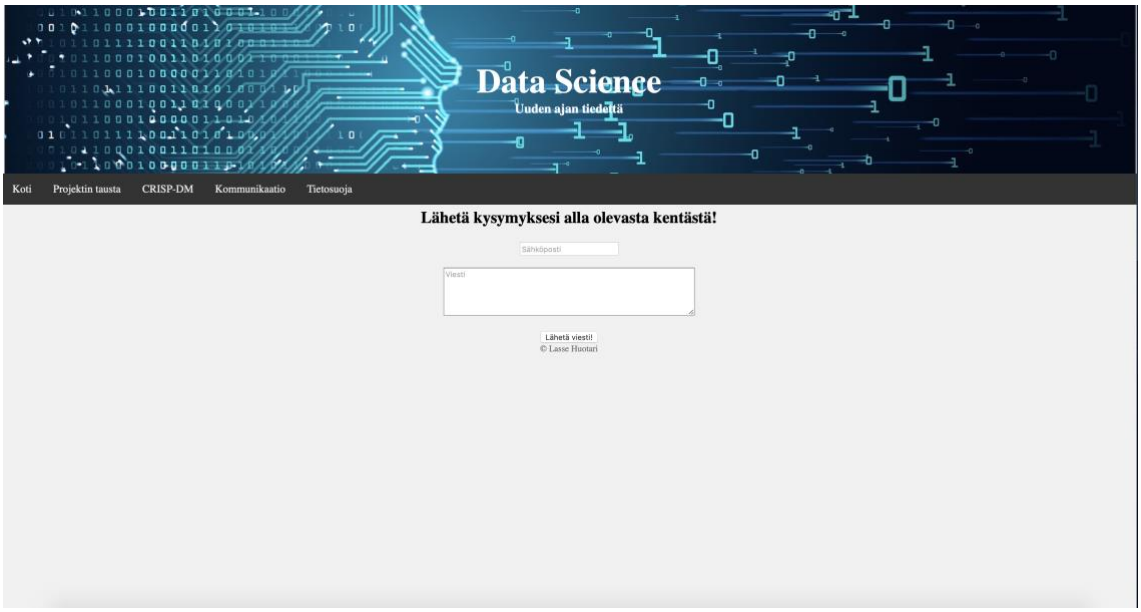


Kuva 7. Tuotteistamisen harjoituksen ohje

Tähän valintaan päädyttiin, koska tehtävissä on tarkoitus käyttää spesifistä työkalua harjoitusten tekemiseen, ja jos työkalu ei ole ennestään tuttu, ei opiskelija koe harjoituksia mielekkääksi, koska työkalun opetteluun menee liikaa aikaa.

5.3.4 Kommunikaatio

Oppimisalustan vaatimuksissa mainittu kommunikaatiotoiminnallisuus on toteutettu viestintäalustana, jossa koulutettava voi lähettää koulutuksen vetäjälle kysymyksen, joka välitetään sähköpostin kautta.



Kuva 8. Oppimisalustan kommunikaatiosivu

5.3.5 Alustan tekninen toteutus

Alusta on rakennettu HTML- ja JavaScript-kielien avulla. Alustan hallinta tapahtuu Githubin välityksellä. Github on Git-pohjainen versionhallintajärjestelmä, joka on esitelty luvussa 4.5 Git. Githubissa säilytetään myös oppimisalustan asiasisällön dokumentteja. Oppimisalusta sijaitsee IBM Cloudissa olevalla web-palvelimella.

Oppimisalustan toteuttamistavaksi valittiin HTML- ja JavaScript-pohjainen nettisivu, koska kyseiset kielet tarjoavat joustavan välineen luoda halutun kaltainen sivuston pohja. Vaihtoehtona olisi ollut käyttää valmiita Open Source -oppimisalustoja, mutta niiden joustavuus ei ollut tarpeeksi suurta alustan luomiseen.

Oppimisalustan asiasisällön hallintaan Github valikoitui tuotteen laajan tuen takia. Github tukee suoraan esimerkiksi markdown-toiminnollisuutta, ja sen näyttämistä

loppukäyttäjälle, joka helpottaa harjoituksissa käytettyjen ohjeiden tekemistä. Lisäksi Github tarjoaa versiohallinnan, jolloin asiasisällön päivittäminen sekä hallinta on helppoa.

6 Ehdotuksen validointi

Tämä osio kattaa tämän tutkimuksen yhteenvedon sekä asiakasyritykselle tehdyn pilotoinnin validoinnin. Lisäksi osio kattaa opinnäytetyön arvioinnin, tuotoksen ja tavoitteiden vertailun sekä arvion opinnäytetyön luotettavuudesta ja oikeellisuudesta.

6.1 Yhteenveto

Tämä tutkimus keskittyi parantamaan asiakasyrityksen koulutusta. Projektin päätavoitteena oli luoda nykyaikainen interaktiivinen oppimisalusta datatieteen ja koneoppimisen koulutukseen. Puhdas itseopiskelu ei tue tarpeeksi datatieteilijöitä materiaalin paljouden ja hajanaisuuden takia. Lisäksi koulutuksen kehittämisellä saadaan liiketoiminnan hyötyjä säästetyllä työajalla, joka olisi mennyt tehottomampaan koulutukseen.

Nykytila-analyysi kuvaa datatieteilijöiden kokemuksia heidän saamastaan koulutuksesta. Lisäksi analyysissä käsiteltiin datatieteilijöiden työssään käyttämiä nykyisiä prosessimalleja ja toimintatapoja. Analyysin lopputuloksena vahvistui käsitys siitä, että datatieteilijät käyttäisivät mieluummin oppimisalustoja, joissa yhdistetään prosessimalli työkaluihin. Nykytila-analyysin tuloksena saatiin myös havainto siitä, että ison organisaation koulutus on usein hajanaista, eivätkä koulutukset kytkeydy yhteen. Toisaalta suuri organisaatio mahdollistaa myös laajan tukiverkoston, jossa mentoroinnin mahdollisuus ja tuen saaminen on helpompaa kuin pienessä organisaatiossa.

Tutkimus ehdottaa ratkaisuksi interaktiivista oppimisalustaa, jossa integroidaan datatietelijöiden yleisesti käyttämä CRISP-DM-prosessimalli ja käytössä olevat työkalut. Oppimisalusta toteutettiin HTML- ja JavaScript-kielten avulla. Oppimisalustan tietosisältö koostuu CRISP-DM-prosessimallin teoriasta, ulkoisista artikkeleista sekä harjoitteiden ohjeista. Teoriasisältö ja ulkoisten artikkeleiden linkit on tallennettu suoraan alustaan,

mutta käytännön harjoitusten ohjeiden tallennuspaikkana toimii Github. Githubilla hallitaan myös alustan versiokehitystä, ja oppimisalustan lähdekoodi on tallennettu sinne. Oppimisalustan vaatimuksissa oleva kommunikoinnin työväline on toteutettu viestialustalla, josta opiskelija voi lähettää sähköpostin suoraan koulutustarjoajan sähköpostiin.

Oppimisalustan harjoitustehtävät koostuvat kahdesta osa-alueesta. Prosessimallin harjoitukset suoritetaan täytettävillä raporteilla, joiden sisällön tuottamiseksi on suunniteltu ja toteutettu erilaisia käytännön harjoituksia. Liiketoiminnan ymmärtämisen tehtävässä simuloidaan asiakkaan kanssa käytyä keskustelua chatbot-palvelun avulla. Muut käytännön tehtävät suoritetaan Watson Studio -palvelussa, joka on datatieteilijän työkalu.

Datan keräämisen aikana haastatteluissa paljastui tarve yhtenäiselle, prosesseja ja työkaluja yhdistelevälle oppimisalustalle. Vastausten mukaan tehokkaimmaksi koulutuksen tarjonnan muodoksi koettiin internetportaali. Tutkimuksessa onnistuttiin rakentamaan oppimisalusta, joka yhdistelee prosessimallien sekä työkalujen koulutuksen tehokkaaksi kokonaisuudeksi.

Oppimisalustan koekäytön jälkeisissä haastatteluissa vastaajat kertoivat kokevansa oppimisalustan tehokkaana tapana kouluttaa itseään datatieteen osaamisalueella. Oppimisalusta koettiin hyödylliseksi, ja sen rakenne tukee työntekijöitä saavuttamaan tavoitteensa koulutuksessa tehokkaasti.

6.2 Oppimisalustan koekäytön palaute

Oppimisalustaa koekäytettiin eri rooleissa olevien toimihenkilöiden toimista. Alustan koekäyttöön osallistui niin datatieteen parissa työskenteleviä henkilöitä kuin täysin ummikkojakin. Koekäytön jälkeen jokaisen käyttäjän kanssa järjestettiin lyhyt palautteenantosessio, jossa haastattelun muodossa käytiin läpi heidän kokemuksiansa ja mielipiteitä oppimisalustan toiminnoista, rakenteesta, soveltuvuudesta sekä hyödyistä.

Ummikkokäyttäjän palautteen mukaan käyttäjä koki oppimisalustan tehokkaaksi ja hyödylliseksi tavaksi yhdistää prosessimallin sekä työkalujen koulutus toisiaan tukevaksi

kokonaisuudeksi. Saadun palautteen perusteella koekäyttäjä koki prosessimallin päälle rakennetun harjoituksen mielekkäänä ja käytännön harjoitukset tarkoituksenmukaisina. Käyttäjän mukaan oppimisalustan harjoitusten suorittaminen antoi hänelle paremman valmiuden työstää datatiedeprojektia omatoimisesti. Palautteen mukaan hän myös koki, että pystyi paremmin selittämään prosessin ja sen eri vaiheet niin kollegoilleen kuin asiakkaallekin. Käyttäjä totesi, että prosessin selitettävyyden takia on oleellista päästä tekemään asioita käytännössä. Jos käytännön harjoituksia ei olisi alustalla tarjolla, hän ei pystyisi selittämään prosessia yhtä hyvin kuin käytännön harjoitukset tehtyään. Sama ilmiö toistui myös toisinpäin käännettäessä: jos oppimisalusta sisältäisi pelkästään käytännön harjoituksia, hän ei kokemansa mukaan pystyisi selittämään prosessia yhtä hyvin kuin silloin, kun käyttäjä oli tutustunut harjoitusten yhteydessä teoriasisältöön. Ummikkokäyttäjän mukaan oppimisalusta nopeuttaisi hänen siirtymisprosessiaan uuteen rooliin datatieteiden pariin. Sama ummikkokäyttäjä myös koki, että oppimisalusta tarjoaa myös hänen osaamistasoaan vastaavaa sisältöä.

Simuloitu keskusteluharjoitus koettiin myös mielekkääksi ja hyödylliseksi. Alustaa käytettäessä käyttäjä olisi toivonut selkeämpää ohjausta kysymyksien asetteluun ja muotoiluun, mutta joutuessaan oikeaan tilanteeseen käyttäjä osaisi käyttää keskustelun perusteella täytettävää liiketoiminnan raporttia pohjana keskustelulle asiakkaan kanssa.

Ummikkokäyttäjän mukaan oppimisalusta toimii paremmin kuin hänen aikaisemmin suorittamat koulutukset ovat toimineet. Harjoitukseen sisällytetty oikeaa elämää simuloiva liiketoiminnan tarina toi harjoitukseen enemmän konkretiaa kuin aikaisemmat hyvin ylätasolla olleet koulutukset. Harjoitukset auttoivat luomaan isomman muistijäljen kustakin prosessin vaiheesta, ja täten paransivat käyttäjän kykyä sisäistä käsitelty aihekokonaisuus. Lisäksi lisäarvoa toi aidon työkalun sisällyttäminen harjoitukseen. Palautteen mukaan alusta myös madaltaa kynnystä lähteä opettelemaan ja kokeilemaan omia datatiedeprojekteja.

Datatieteen saralla kokeneen käyttäjän palautteen mukaan oppimisalustan toimii hyvin osana koulutusta. Erityisesti teorian yhdistäminen käytännön harjoituksiin toi käyttäjän mukaan hyvää vaihtelua jo olemissa oleviin joko-tai-harjoituksiin ja koulutukseen. Käyttäjän palautteen mukaan kuitenkin alustassa on vielä kehitettävää, esimerkiksi asettelun ja ulkoasun osalta. Käyttäjä koki jossain määrin alustan käytön tehottomaksi, koska alustan tietosisältö on hyvin tekstipainotteista. Parannukseksi tähän käyttäjä

ehdotti erilaisten visualisointien lisäämistä alustaan ja tietosisällön muuttamista osittain visuaaliseen muotoon. Lisäksi kokenut käyttäjä kaipasi ummikkokäyttäjää enemmän aihekokonaisuuksien ja eri termien syvempää läpikäyntiä. Esimerkiksi datan ymmärtämisen vaiheessa esiintyvien erilaisten korrelaatiomallien syvempi avaaminen olisi ollut hyvä lisäominaisuus.

Käyttäjä koki myös, että tehtävien vapaavalintaisuus ja itseohjautuvuus ei tuo tarpeeksi esille koulutuksen pakollisuutta. Parannusehdotuksena käyttäjä mainitsi esimerkiksi raporttien vertaisarvioinnin. Kuitenkin jo nyt alusta tarjoaa hyvän pohjan koulutukseen ja lisäksi käyttäjä koki alustan sovellettavuuden hyvänä. Palautteen mukaan alustaa voidaan käyttää asiakasprojektien suorittamiseen, ja se tuo tarvittavat valmiudet käydä projektia läpi asiakkaan kanssa. Erityisen hyvänä ominaisuutena palautteessa mainittiin itsestään selvien asioiden dokumentoinnin. Kokeneen käyttäjän mielestä itse prosessin selitettävyys kollegalle ja asiakkaalle on parempi oppimisalustan koulutuksen suoritettuaan.

Chatbot-toiminnallisuuden datatieteen parissa työskennellyt koki hyvänä yhden kerran harjoituksena, mutta jotta siitä olisi hänelle todellista hyötyä chatbotin tulisi olla huomattavasti moniuloitteisempi, sekä sillä tulisi olla kyky eläytyä erilaisiin rooleihin. Käyttäjä kuitenkin koki, että oppimisalusta voisi olla tehokas keino kouluttautua, kunhan edellä mainitut parannusehdotukset on toteutettu.

Kokonaisuutena oppimisalustaa voidaan pitää palautteen perusteella tehokkaana ratkaisuna osana muuta koulutuskokonaisuutta, mutta nykytilassaan se palvelee ennemmin uuteen rooliin valmistautuvaa tai omaa osaamisaluettaan kasvattavaa kohdeyleisöä. Mahdollisen jatkokehityksen osalta alusta voi kuitenkin myös toimia jatkuvan koulutuksen tarjoajana kokeneelle datatieteilijälle.

6.3 Opinnäytetyön arviointi

Tutkimuksen aikajänne kasvoi merkittävästi alkuperäisestä suunnitelmasta, mutta samaan aikaa työn merkittävyys ja tuotos kasvoi. Opinnäytetyön aihealueiden parissa tehtävä päivätyö helpotti työn tekemistä, mutta myös nosti työlle asetettuja kriteereitä. Ohjaavan opettajan tuki helpotti merkittävästi tutkimuksen raportointia sekä auttoi

asioiden muovaamista raportin muotoon. Asiakasyrityksen vastaanottavaisuus ja toimihenkilöiden avuliaisuus helpottivat merkittävästi oppimisalustan koekäyttöä ja suunnittelua. Haastatteluista saatiin paljon tutkimuksen kannalta oleellista taustatietoa, jota olisi ollut lähes mahdotonta saada vain yleisiä tietolähteitä käyttäen. Tämä tuki mahdollisti onnistuneen projektin ja sen hyvät lopputulokset. Asiakasyritykseltä saatu palaute oppimisalustasta oli positiivista, ja tutkimus koettiin merkitykselliseksi asiakkaan toimesta. Oppimisalustaa ei voida pitää uutena keksintönä, mutta siihen lisätyt uudet komponentit ja toiminnallisuudet toivat uudella tavalla yhteen jo olemassa olevat koulutuskokonaisuudet. Koekäyttöön osallistuneiden määrä olisi voinut olla isompi, mutta heidän eri taustansa mahdollistivat luotettavan ja todellisen kuvan alustan sopivuudesta asiakasyrityksen käyttöön.

Projektin alussa asiakasyrityksen kanssa määriteltiin tavoitteeksi parantaa toimihenkilöiden ja työntekijöiden itsekoulutuksen mahdollisuuksia luomalla konsepti validoidusta, interaktiivisesta oppimisalustasta, jossa yhdistyvät datatieteelle ominaiset prosessimallit sekä käytännön työkalut.

Projektin aikana kehitettiin toiminnassa oleva oppimisalusta, joka sisältää alalla tyypilliseen prosessimalliin pohjautuvan, käytännönharjoituksilla täydennetyn opintokokonaisuuden. Täten voidaan todeta, että projektilla kyettiin vastaamaan asetettuihin tavoitteisiin.

Lähteet

Mikä Datatiede? Verkkoaineisto. Tampereen teknillinen yliopisto. <https://www.datatiede.fi/mika-datatiede/> Luettu 1.9.2019.

Koneoppiminen. Verkkoaineisto. ite wiki. <https://www.itewiki.fi/opas/koneoppiminen/> Luettu 1.9.2019.

Vorhies, William. 2016. CRISP-DM – a Standard Methodology to Ensure a Good Outcome. Verkkoaineisto. Data Science Central. <https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome> 26.6.2016. Luettu 1.9.2019

Lo, Frank. What is Data Science?. Verkkoaineisto. Datajobs. <https://datajobs.com/what-is-data-science> Luettu 1.9.2019.

Warren, Cameron. 2019. Don’t Do Data Science, Solve Business Problems. Verkkoaineisto. Towards Data Science<https://towardsdatascience.com/dont-do-data-science-solve-business-problems-6b70c4ee0083> 6.3.2019. Luettu 1.9.2019.

Bowne-Andersson, Hugo. 2018. What Data Scientist Really Do, According to 35 Data Scientists. Verkkoaineisto. Harvard Business Review. <https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists> 15.8.2018. Luettu 1.9.2019

Ray, Tanmoy. 2018. Data Engineer vs Data Scientist – Background, Responsibilities, Skills, and Job Prospects. Verkkoaineisto. Stoodnt. <https://www.stoodnt.com/blog/data-engineer-vs-data-scientist/> 10.8.2018. Luettu 1.9.2019.

Pickell, Devin. 2018. Structured vs Unstructured Data – What's the Difference? Verkkoaineisto. g2. <https://learn.g2.com/structured-vs-unstructured-data> 16.11.2018. Luettu 1.9.2019.

Naur, Peter. 1974. Concise Survey of Computer Methods. Lund, Sweden, Studentlitteratur.

Data Science and Its Growing Importance. Verkkoaineisto. Educba. <https://www.educba.com/data-science-and-its-growing-importance/> Luettu 1.9.2019.

Twin, Alexandra. 2019. Data Mining. Verkkoaineisto. Investopedia. <https://www.investopedia.com/terms/d/datamining.asp> 18.8.2019. Luettu 1.9.2019.

CRISP-DM Help Overview. Verkkoaineisto. IBM. <https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.html> Luettu 17.11.2018

What is the CRISP-DM methodology? Verkkoaineisto. Smart Vision Europe. <https://www.sv-europe.com/crisp-dm-methodology/> Luettu 17.11.2018.

Chatbot: What is Chatbot? Why are Chatbots Important? Verkkoaineisto. Expert System. <https://www.expertsystem.com/chatbot/> Luettu 1.9.2019

Combining Search, Chatbots, and Question Answering to Deliver Holistic Enterprise Knowledge. Verkkoaineisto. Search Technologies. <https://www.searchtechnologies.com/blog/search-chatbot-question-answering> Luettu 1.9.2019

What Is Git. Verkkoaineisto. Atlassian <https://www.atlassian.com/git/tutorials/what-is-git> Luettu 1.9.2019.

Piotrowski, Michael. What is an E-Learning Platform? Verkkoaineisto. ZHAW Zurich University of Applied Sciences, Switzerland. <http://www.irma-international.org/viewtitle/43445/> Luettu 2.9.2019.

Laithangbam, Michael. 2017. What is an LSM? Components, Features, Deployment Types, Users and More. Verkkoaineisto. ProProfs. <https://www.proprofs.com/c/lms/what-is-an-lms/> 1.6.2017. Luettu 2.9.2019.