ROBOCOAST EDIH European Digital Innovation Hubs Network Prizztech SATAKUNTALIITTO Regional Council of Satakunta Leverage from the EU 2014–2020 Euroopan unioni Euroopan aluekehitysrahasto

3D Data Economy Satakunta's Success Factor in Digital Green Growth

Pilotti 1

Datan hyödyntämisen konsepti teollisuus- ja automaatioympäristössä

Hanke on rahoitettu REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Hankkeen *3D Data Economy Satakunta's Success Factor in Digital Green Growth* yrityspilotti: Case Cimcorp Ulvila

Raportti

Jari Turunen & Tarmo Lipping, Tampereen yliopisto

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Sisällysluettelo

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Kuvaluettelo

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Tiivistelmä

Tässä raportissa esitellään **3D Data Economy Satakunta's Success Factor in Digital Green Growth** -hankkeen yrityspilotti, joka toteutettiin yhteistyössä Cimcorp Oy Ulvilan kanssa. Pilotissa tuotettiin datan analyysiin tarkoitettua koodia sekä kehitettiin analyysikonsepti robotiikkajärjestelmän kapasiteetin tarkastelua varten yhteistyössä Cimcorp Oy:n asiantuntijoiden kanssa. Raportissa kuvataan ensin pilottikohde eli Cimcorp Oy:n varastonhallintaan tarkoitettu robottijärjestelmä. Sen jälkeen esitellään pilotin tavoitteet ja kuvataan yleisluontoisesti käytettävissä ollut data sekä toimenpiteet, joita tarvittiin datan esikäsittelyyn ennen mallinnusta. Pilotin keskeinen toimenpide oli kapasiteetin tarkasteluun tarvittavien eri mallinnusmenetelmien kehittäminen, vertailu ja validointi selvittääkseen robottijärjestelmän eri tehtäviin kuluvan ajan riippuvuutta tehtävien ominaisuuksista. Raportissa kuvataan yleisluontoisesti pilotissa testattuja mallinnusmenetelmiä, niiden ominaisuuksia sekä niiden validoinnissa käytettyjä työkaluja ja menetelmiä. Lopuksi esitellään tyypillisen data-analytiikkaprojektin prosessikaavio sekä tuodaan esiin tärkeimpiä havaintoja pilotista.

1. Johdanto

Datan kerääminen teollisuusympäristössä on yleistynyt merkittävästi viimeisen vuosikymmenen aikana. Datan hyödyntämisen avulla pyritään optimoimaan toimintaa, vähentämään kustannuksia, nostamaan toimintavarmuutta sekä luomaan uutta liiketoimintaa. Kerätyn datan hyödyntäminen ei kuitenkaan ole aina suoraviivaista ja vaatii osaamista mm. tilastollisista menetelmistä ja koneoppimisesta sekä sovelluskohteen perusteellista tuntemusta.

Tässä raportissa kuvataan vuoden 2023 alkupuolella toteutetun data-analytiikkapilotin toteutusta ja tuloksia. Pilottikohteena oli Cimcorp Oy:n robottiratkaisu varastonhallintaan, mutta pilotissa käytetyt menetelmät ja ratkaisut ovat yleistettävissä laajemmin vastaaviin sovelluskohteisiin. Raportissa kuvataan ensin pilottikohde, jonka jälkeen kerrotaan pilotin tavoitteet. Luvussa 4 annetaan yleiskuva käytettävissä olevasta datasta ja sen esikäsittelymenetelmistä. Tämän jälkeen kerrotaan data-analytiikkaratkaisuissa käytettävistä työkaluista. Luvussa 6 kuvataan pilotissa käytettyjä malleja, joiden avulla voidaan tarkastella eri muuttujien vaikutusta kohdemuuttujaan tarkoituksena selvittää, mistä kohdemuuttujan ei-toivottu käyttäytyminen johtuu. Lopuksi yhteenveto-osuudessa kuvataan data-analytiikkapilotin yleinen konsepti automaatio- ja robotiikkaympäristössä sekä vedetään pilotista opittu yhteen.

2. Pilottikohteen kuvaus

Cimcorp Oy valmistaa varastonhallinta-automatiikkaa asiakkailleen. Varastonhallinta koostuu kolmiulotteisessa koordinaatistossa toimivista roboteista, kuljettimista sekä ohjausjärjestelmistä. Esimerkki tällaisesta järjestelmästä on Kuvassa 1.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Kuva 1. *Tuoretavaravarasto © First Production Marius Tikkanen, Cimcorp Media Hub [1]*

Kuvan 1 kaksi robottia täydentävät lattiavarastoa sisääntulokuljettimilta tulevilla laatikkopinoilla. Varastonhallintajärjestelmä pitää kirjaa tuotteiden paikoista ja samalla tuotteiden ikääntymisestä, joten varastojärjestelmän asiakkaat saavat aina tuoreusasteeltaan erinomaisia tuotteita. Varastot ovat nopeita välivarastoja tuoretuotteiden kohdalla, joten tilauksen tultua varastorobotit hakevat oikeat laatikot ja siirtävät ne ulosmenokuljettimille pakkausta ja kuljetusta varten. Varastorobotit toimivat autonomisesti ja keskeytykset johtuvat pääasiassa varaston siivouksesta ja/tai laitehuolloista.

3. Pilotin tavoitteet

Pilotin tavoitteena oli selvittää, miten teollisuusympäristöstä kerätyn data pohjalta voidaan parhaiten selvittää kriittisiin kohdemuuttujiin (esim. suoritusaikaa tai laatua kuvaavat muuttujat) eniten vaikuttavat tekijät. Pilotin alkuvaiheessa oli tarkoitus käydä yhdessä pilottikohteen edustajan kanssa läpi kerätty data ja määritellä siitä ennakkotiedon pohjalta avainmuuttujat. Tarkoitus oli myös määritellä poikkeamat tai kohdemuuttujien ei-toivotut tilat, joita datan pohjalta lähdetään selvittämään. Valittuja muuttujia oli tarkoitus tarkastella visualisointien avulla ja valita data-analytiikkaan tarkoituksenmukaiset työkalut.

Pilotin seuraavassa vaiheessa oli tavoitteena sovittaa kerättyyn dataan simulointimalleja, joiden avulla voidaan selvittää avainmuuttujien ja esim. tuotantolinjan kapasiteettia kuvaavien mittareiden välisiä riippuvuuksia. Tarkastelemalla näitä riippuvuuksia voidaan selvittää, miten ei-toivotut tilanteet syntyvät ja miten niitä voidaan karsia. Pilotissa oli myös tarkoitus käyttää kehitettyjä malleja generatiivisesti, eli tarkastella mallien tuottamien kohdemuuttujien käyttäytymistä, kun avainmuuttujia vaihdellaan tilastollisesti tiettyjen sääntöjen mukaisesti.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Pilotin tuloksena konseptoitiin teollisuusympäristössä toteutettavan data-analytiikan kehittämisprosessi. Konsepti ottaa kantaa tarvittavan datan laatuun, käytettäviin työkaluihin ja malleihin sekä tulosten tulkintaan.

4. Datan esikäsittely

Cimcorp Oy:n vuositasolla keräämä data käsittää miljoonia rivejä yksittäistä robottisolua kohden. Roboteista kerätään varastotietoa (esimerkiksi laatikkopinojen korkeus) sekä robotin toimintaan liittyvää tietoa (esimerkiksi kulkunopeudet ja -matkat x- ja y-suunnissa).

4.1 Keskiarvotettu data

Datan tarkastelussa lähdettiin liikkeelle keskiarvotetusta datasta, jossa jokainen rivi kertoo 6 minuutin aikajakson tapahtumista (Kuva 2). Datan avulla oli tarkoitus selvittää, millaisista tekijöistä robottijärjestelmän kapasiteetti riippuu. Kellonajan ja päivämäärän jälkeen aineistossa on sarakkeita, jotka ilmaisevat esim. robotin tehokkaan työskentelyajan tai eri tehtävien osuuden k.o. jaksossa. Yhteensä aineisto sisälsi yli 150 muuttujaa. Pian kuitenkin huomattiin, että syy-seuraussuhteita ei ollut mahdollista tarkastella keskiarvotetusta datasta halutulla tarkkuudella. Nopeita tapahtumia voi olla useita 6 minuutin aikana ja niiden yksityiskohtaisiin ominaisuuksiin ei keskiarvoilla pääse käsiksi. Jos pidempikestoinen tapahtuma osuu kahden 6 minuutin jakson väliin, se kirjautuu jommankumman jakson tapahtumaksi, jolloin tieto vääristyy. Samaan jaksoon osuu myös erityyppisiä tehtäviä, jolloin tehtävätyyppien tarkastelu erillisinä ei ole mahdollista.

Keskiarvotetun datan tapauksessa jakson pituus ratkaisee, kuinka yksityiskohtaisesti datasta päästään tarkastelemaan robotin toimintaa. Pilotissa testattiin datan hyödynnettävyyttä jakson pituutta muuttamalla (3 min, 6 min, 12 min ja 30 min). Tämä tarkoitti käytännössä keskiarvotetun datan uudelleenluontia alkuperäisestä tietokannasta. Jakson pituuden valinta vaikuttaa keskiarvodatan rivien määrään; mitä pidempi jakso on, sitä vähemmän tulee aineistoon rivejä. Pidempi keskiarvotusjakso vähentää jaksorajojen aiheuttamia vääristymiä, mutta samalla yhä enemmän yksityiskohtia robotin toiminnasta jää piiloon.

Alkuselvittelyn jälkeen päätettiin hylätä keskiarvotetun datan käyttö ja siirryttiin tarkastelemaan tehtäväkohtaista dataa.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1																					
2	13.7.2021 10:12	12	1.7.2021	2021-28	Tue	1	1	0.0	1	1	0.0	0.0	505.0	1	0.0	0.0	0.0	0.0	1.0	0.0	0.14515833333333336
3	13.7.2021 10:18	12	1.7.2021	2021-28	Tue	3	3	0.09130555555555554	3	3	0.0	0.0	502.0	3	0.0	0.0	0.0	0.0	1.0	0.0	0.42345
4	13.7.2021 10:24	12	1.7.2021	2021-28	Tue	3	3	0.06640138888888889	3	3	0.0	0.0	499.0	3	0.0	0.0	0.0	0.0	1.0	0.0	0.4150166666666667
5	13.7.2021 10:30	12	1.7.2021	2021-28	Tue	6	6	0.12496527777777777	6	6	0.0	0.0	495.0	6	0.0	0.0	0.0	0.0	1.0	0.0	0.7945416666666667
6	13.7.2021 10:36	12	1.7.2021	2021-28	Tue	8	8	0.16890416666666666	8	8	0.0	0.0	491.0	8	0.0	0.0	0.0	0.0	1.0	0.0	1.0573166666666667
7	13.7.2021 10:54	12	1.7.2021	2021-28	Tue	3	2	0.043704166666666676	3	2	0.0	0.0	490.0	2	0.0	0.0	0.0	0.0	1.0	0.0	0.2652416666666667
8	13.7.2021 11:00	12	1.7.2021	2021-28	Tue	1	2	0.041063888888888886	1	2	0.0	0.0	489.0	2	0.0	0.0	0.0	0.0	1.0	0.0	0.2636833333333334
9	13.7.2021 11:48	12	1.7.2021	2021-28	Tue	1	0	0.0200875	0	0	1.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	13.7.2021 11:54	12	1.7.2021	2021-28	Tue	12	13	0.5033180555555555	2	2	10.0	11.0	488.0	15	0.5333333	0.0	0.1333333	0.2	0.1333333333	0.0	1.3905416666666668
11	13.7.2021 12:00	12	1.7.2021	2021-28	Tue	20	14	0.44580972222222226	1	1	19.0	13.0	488.0	19	0.3157894	0.0	0.1578947	0.4736842105	0.0526315789	0.0	1.9228
12	13.7.2021 12:06	12	1.7.2021	2021-28	Tue	3	9	0.1645416666666667	1	1	2.0	8.0	489.0	8	0.5	0.0	0.125	0.25	0.125	0.0	0.7189
13	13.7.2021 12:12	12	1.7.2021	2021-28	Tue	17	17	0.4515180555555556	2	2	15.0	15.0	488.0	26	0.2307692	0.0	0.3461538	0.3461538461	0.0769230769	0.0	2.1447000000000003
14	13.7.2021 12:18	12	1.7.2021	2021-28	Tue	42	40	1.668320833333333	6	6	36.0	34.0	487.0	45	0.4666666	0.0	0.1111111	0.2888888888	0.1333333333	0.0	4.2984333333333336
15	13.7.2021 12:24	12	1.7.2021	2021-28	Tue	1	2	0.025291666666666667	0	0	1.0	2.0	0.0	2	1.0	0.0	0.0	0.0	0.0	0.0	0.1610666666666667
16	13.7.2021 12:30	12	1.7.2021	2021-28	Tue	10	11	0.26102916666666665	0	0	10.0	11.0	493.0	18	0.0	0.0	0.3888888	0.6111111111	0.0	0.0	1.6698583333333334
17	13.7.2021 13:00	12	1.7.2021	2021-28	Tue	4	4	0.09735416666666669	0	0	4.0	4.0	498.0	8	0.0	0.0	0.5	0.5	0.0	0.0	0.6123999999999999
18	13.7.2021 13:42	12	1.7.2021	2021-28	Tue	2	2	0.13581944444444444	0	0	2.0	2.0	502.0	4	0.0	0.0	0.5	0.5	0.0	0.0	0.28478333333333333
19	13.7.2021 15:06	12	1.7.2021	2021-28	Tue	0	0	0.0	0	0	0.0	0.0	0.0	2	0.0	0.0	1.0	0.0	0.0	0.0	0.07531666666666667
20	14.7.2021 11:06	12	1.7.2021	2021-28	Wed	4	3	0.4441416666666667	4	3	0.0	0.0	500.0	4	0.0	0.0	0.0	0.0	1.0	0.0	0.4830750000000001
21	14.7.2021 11:12	12	1.7.2021	2021-28	Wed	3	2	0.9734944444444446	3	2	0.0	0.0	498.0	2	0.0	0.0	0.0	0.0	1.0	0.0	0.7638750000000001
22	14.7.2021 11:18	12	1.7.2021	2021-28	Wed	3	2	1.1786819444444445	3	2	0.0	0.0	496.0	2	0.0	0.0	0.0	0.0	1.0	0.0	0.6610916666666667
23	14.7.2021 11:24	12	1.7.2021	2021-28	Wed	10	11	1.361075	10	11	0.0	0.0	490.0	11	0.0	0.0	0.0	0.0	1.0	0.0	2.212775
24	14.7.2021 11:30	12	1.7.2021	2021-28	Wed	0	0	1.0	0	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25	14.7.2021 11:36	12	1.7.2021	2021-28	Wed	0	0	1.0	0	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	14.7.2021 11:42	12	1.7.2021	2021-28	Wed	0	0	1.0	0	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

***Kuva 2**. Ote keskiarvodatasta. Jokainen rivi käsittää 6 minuutin aikajaksoa robotin toiminnassa.*

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

4.2 Tehtäväkohtainen data

Ote tehtäväkohtaisesta lokidatasta, johon robottien tapahtumat tallentuvat lähes reaaliaikaisesti, on Kuvassa 3. Tehtäväkohtaiset muuttujat poikkeavat keskiarvodatassa olevista muuttujista ja kuvaavat robotin toimintaa kyseisen tehtävän yhteydessä. Aineiston käsittelyn yhteydessä muuttujien joukkoa vaihdeltiin sen mukaan, millaiset muuttujat arvioitiin vaikuttavan tutkittavaan ilmiöön (tässä tapauksessa tehtävän suoritusaikaan). Tarkasteltavien muuttujien määrä vaihteli 40…50 paikkeilla. Pilotissa käsiteltiin n. vuoden ajalta kerättyä dataa, joka käsitti noin 6,5 miljoonaa riviä. Robottisolussa esiintyy kahdeksan erityyppistä tehtävää, joiden määrä vaihteli pilottidatassa muutamasta tuhannesta muutamaan miljoonaan (Taulukko 1). Myös eri muuttujien merkitys eri tehtävien yhteydessä vaihtelee; esimerkiksi osassa tehtävistä robotti ei liiku, jolloin liikettä kuvaavat muuttujat ovat tyhjiä.

Sarake1	Sarake2	Sarake3	Sarake4	Sarake5	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake	Sarake
8.2.2022 18:57	12+C20	ROB002	101 10.15	0	15	2	1	1	83	89	0	600	0	0	98	0	0	0	38	15	-6	10665	9313	18847
8.2.2022 19:12	13+C20	ROB001	101 8.863	0	14	0	1	1	134	149	0	600	0	0	149	0	0	0	36	8	-13	7365	8528	835
8.2.2022 19:30	13+C20	ROB002	101 9.224	0	16	0	3	3	265	275	0	600	0	0	280	0	0	0	37	12	-15	11765	7748	18847
8.2.2022 20:04	12+C20	ROB001	101 9.74	0	18	2	3	3	265	265	0	600	0	0	280	0	0	0	39	14	-17	8465	10093	835
8.2.2022 20:04	12+C20	ROB002	101 6.506	0	16	2	10	10	904	910	0	600	0	0	918	0	0	0	13	1	-4	11765	8533	10665
8.2.2022 20:04	12+C20	ROB002	101 6.695	0	17	2	2	2	221	221	0	600	0	0	235	0	0	0	14	3	-4	11765	10093	10665
9.2.2022 4:48	12+C20	ROB001	101 8.334	0	18	2	5	5	561	555	0	600	0	0	576	0	0	0	31	11	-13	1385	6973	6915
9.2.2022 4:48	12+C20	ROB002	101 8.003	0	19	2	5	5	561	563	0	600	0	0	576	0	0	0	29	9	-4	17747	7753	12765
9.2.2022 4:49	12+C20	ROB001	101 8.616	0	20	2	2	2	221	215	0	600	0	0	235	0	0	0	28	9	-8	1935	5413	6915
9.2.2022 4:49	12+C20	ROB002	101 8.076	0	17	2	5	5	561	563	0	600	0	0	576	0	0	0	30	8	-7	17747	5413	12765
9.2.2022 4:50	12+C20	ROB001	101 8.325	0	20	2	11	11	1242	1232	0	600	0	0	1257	0	0	0	27	4	-8	2965	5413	7365
9.2.2022 4:50	12+C20	ROB001	101 7.71	0	22	2	8	8	1316	1312	0	600	0	0	1330	0	0	0	25	9	-10	3515	11653	6915
9.2.2022 4:50	12+C20	ROB002	101 7.923	0	17	2	5	5	820	824	0	600	0	0	834	0	0	0	28	9	-11	17197	6193	12765
9.2.2022 4:50	12+C20	ROB001	101 7.77	0	22	2	7	7	630	629	0	600	0	0	645	0	0	0	26	6	-9	3515	6973	7465
9.2.2022 4:51	12+C20	ROB002	101 7.823	0	16	2	10	10	904	911	0	600	0	0	918	0	0	0	26	6	-5	16165	6973	12215
9.2.2022 5:21	12+C20	ROB001	101 7.867	0	15	1	11	11	2210	2201	0	600	0	2544	2225	448	0	336	25	5	-5	5525	2180	1935
9.2.2022 5:27	12+C20	ROB002	101 8.907	0	15	2	15	15	2136	2134	0	600	0	0	2151	0	0	0	36	12	-12	11765	10873	18847
9.2.2022 5:27	12+C20	ROB002	101 6.195	0	16	2	10	10	904	909	0	600	0	0	918	0	0	0	8	1	-1	12315	6973	11765
9.2.2022 5:27	12+C20	ROB001	101 9.142	0	15	1	10	10	2009	1995	0	600	0	0	2023	0	0	0	29	7	-5	5525	2180	835
9.2.2022 5:29	12+C20	ROB002	101 8.047	0	15	1	10	10	2009	2001	0	600	0	0	2023	0	0	0	18	2	-3	14010	2180	11765
9.2.2022 5:29	12+C20	ROB001	101 8.684	0	15	1	10	10	2009	1998	0	600	0	0	2023	0	0	0	25	7	-7	5525	2180	1935
9.2.2022 5:30	12+C20	ROB002	101 8.648	0	15	1	10	10	2009	2002	0	600	0	0	2023	0	0	0	23	4	-2	14010	2180	17747
9.2.2022 5:32	12+C20	ROB001	101 8.436	0	15	1	15	15	1696	1685	0	600	0	0	1711	0	0	0	20	6	-2	5525	2180	2965
9.2.2022 5:32	12+C20	ROB002	101 7.864	0	16	0	1	1	134	139	0	600	0	0	149	0	0	0	29	5	-10	12315	7753	17197
9.2.2022 5:34	12+C20	ROB002	101 8.544	0	15	1	15	15	1696	1689	0	600	0	0	1711	0	0	0	21	12	-1	14010	2180	17197
9.2.2022 5:36	12+C20	ROB001	101 7.917	0	19	2	6	6	849	846	0	600	0	0	864	0	0	0	26	10	-4	7365	6973	3515
9.2.2022 5:36	12+C20	ROB002	101 7.989	0	16	2	10	10	1421	1420	0	600	0	0	1436	0	0	0	25	9	-4	12315	8533	16165
9.2.2022 5:36	12+C20	ROB001	101 8.034	0	18	2	15	15	1696	1686	0	600	0	0	1711	0	0	0	21	5	-6	4615	6193	7365
9.2.2022 5:49	12+C20	ROB001	101 7.896	0	15	1	15	15	2136	2119	0	600	0	0	2151	0	0	0	16	1	-3	5525	2180	7365
9.2.2022 5:52	12+C20	ROB002	101 7.823	0	15	1	13	13	2142	2140	0	600	0	0	2156	0	0	0	18	0	-2	14010	2180	11765
9.2.2022 5:57	12+C20	ROB001	101 6.938	0	20	2	5	5	1131	1127	0	600	0	0	1145	0	0	0	16	4	-5	4615	6973	6265

Kuva 3. *Ote tehtäväkohtaisesta datasta. Rivien aikaleimat ovat tehtäväkohtaisia eli rivit käsittävät eripituisia tehtäviä. Myös muuttujat ovat tehtäväkohtaisia.*

Taulukko 1. *Robottisolun tehtävien määrät pilottidatassa*

Tehtävä 1	3077011
Tehtävä 2	11153
Tehtävä 3	442441
Tehtävä 4	1392
Tehtävä 5	138960
Tehtävä 6	204679
Tehtävä 7	2051788
Tehtävä 8	581774

4.3 Datan esikäsittely

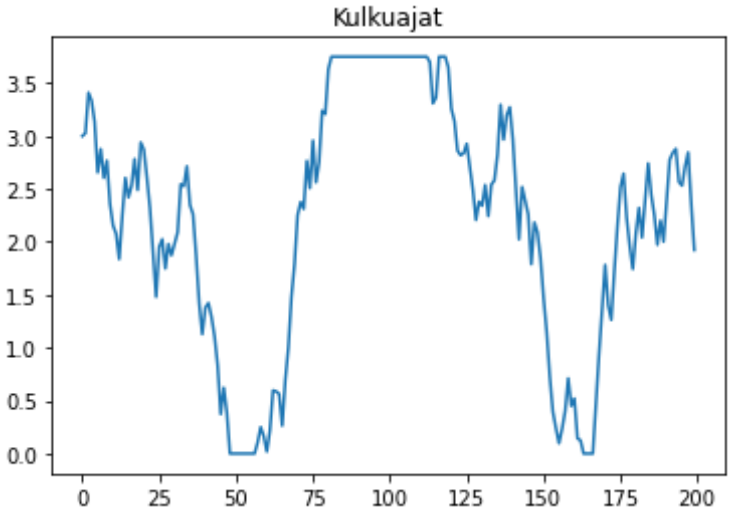
Datan puhdistaminen ennen analyysia on yleensä välttämätön toimenpide. Puhdistuksen tekee yleensä data-analyytikko, mutta se on tehtävä yhteistyössä pilottikohteen asiantuntijan kanssa. Kuvassa 3 on esimerkki datasta, mikä vaatii toimenpiteitä. Yleisimpiä esikäsittelyn operaatioita ovat puuttuvat tai merkityksettömän datan korvaaminen tai poistaminen sekä virheellisten arvojen havaitseminen ja poistaminen.

Osa aineiston soluista saattavat olla tyhjiä tai sisältävät pelkkiä nollia. Kun data luetaan käsittely-ympäristöön (esim. python-ympäristö), tyhjät solut saavat yleensä arvon NaN (Not-A-Number). Joskus

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

datassa itsessään voi olla soluja, joiden arvo on määrittelemätön, jolloin jo lähtödatassa voi esim. olla merkintä NA (Not Applicable). Yleispätevää sääntöä sille, miten puuttuvan datan osalta on toimittava, ei ole mahdollista antaa, vaan toimintatapa riippuu kyseessä olevan muuttujan tyypistä ja merkityksestä. Jos muuttuja käsittää aikasarjaa (esim. lämpötila) ja datasta puuttuu yksittäisiä arvoja, puuttuvat arvot voidaan usein korvata interpoloimalla. Jos muuttuja on kategorinen (esim. tehtävän tyyppi tai robotin käsittelemän tavaran koko), puuttuvaa data ei yleensä voida korvata. Silloin datan mallinnusmenetelmästä ja -algoritmista riippuu, voidaanko data syöttää mallille puuttuvine soluineen (esim. syväoppivan neuroverkon tapauksessa) tai kannattaako datasta poistaa rivit, jotka sisältävät puuttuvia soluja. Jos tietty muuttuja sisältää kokonaan NaN-arvoja, se kannattaa poistaa aineistosta. Samoin esimerkiksi kokonaan nollia käsittävä muuttuja kannattaa poistaa, koska sillä ei ole mallinnuksen kannalta lisäarvoa. Puuttuvan datan käsittelyyn vaikuttaa myös se, käsitelläänkö datassa olevia rivejä sekvenssinä, jonka alkiot riippuvat toisistaan, vai ovatko ne riippumattomia näytteitä robotin toiminnasta. Jos dataa käsitellään sekvenssinä, rivien poistaminen rikkoo sekvenssin ja voi vaikuttaa näin ollen analyysin tulokseen. Pilotin yhteydessä käytettiin erilaisia tapoja käsitellä puuttuvaa data riippuen muuttujan tyypistä ja analyysin luonteesta.

Kun puuttuvat arvot ovat yleensä aina tiedossa, virheellisten arvojen havaitseminen voi olla joskus vaikea. Virheelliset arvot voivat johtua esim. anturin vioittumisesta (yleensä aikasarjadatan yhteydessä), virheellisistä merkinnöistä (jos esimerkiksi osa datasta kirjataan manuaalisesti) tai virheistä datankeruujärjestelmässä (esimerkiksi data kirjautuu väärään formaattiin tai synkronisaatio muuttujien välillä ei toimi tai muuta vastaavaa). Virheellisten arvojen havainnoinnissa on avuksi muuttujien tarkastelu visualisointityökaluilla. Erityyppisten muuttujien visualisoinnista on esimerkkejä Kuvassa 6. Aikasarjan tapauksessa muuttuja voidaan esittää jatkuvana viivana, kuten Kuvassa 4. Kuva esittää muuttujaa, jossa esiintyy virheellisiä arvoja ainakin välillä 75…120 saturoinnista johtuen. Saturoinnin syy voi olla anturin väärä kalibrointi, mutta esimerkiksi myös tapaus, jossa muuttujalle varattu bittimäärä järjestelmän muistissa ei riitä muuttujan koko vaihteluvälin esittämistä. Saturoitumistapauksissa esikäsittelytoimenpiteistä on päätettävä sovelluskohteen asiantuntijan kanssa. Pilottikohteen datassa ei havaittu merkittävästi virheellisiä arvoja. Tarkasteltavat muuttujan olivat lähinnä kategorisia muuttujia, jolloin saturaatio-ongelmaa ei esiintynyt.



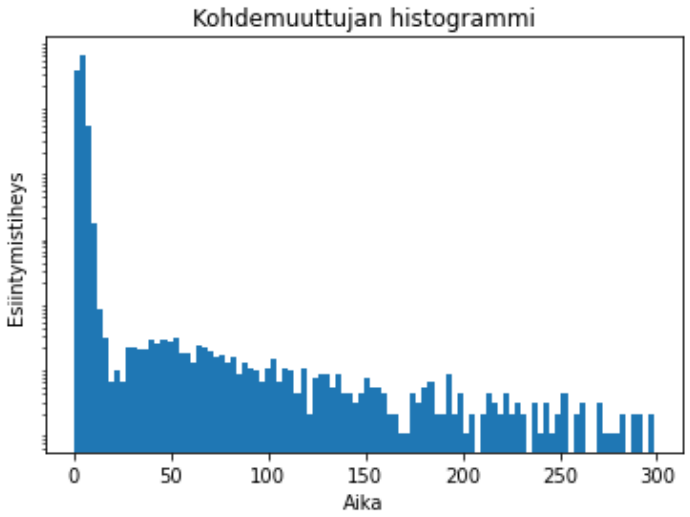
Kuva 4. *Esimerkki saturoitumisesta.*

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

4.4 Datan visualisointi

Edellisessä kappaleessa käsiteltiin, miten muuttujien visualisoinnin avulla voidaan havaita virheellisiä arvoja datassa. Pilotissa tarkoituksena oli löytää muuttujat, jotka selittävät kohdemuuttujan (tässä tapauksessa tehtävän suorittamiseen kuluvan ajan) vaihtelun. Tällöin aineistoa voidaan visualisoida piirtämällä kohdemuuttujan histogrammi sekä muuttujien välisiä korrelaatioita.

Esimerkki kohdemuuttujan histogrammista on Kuvassa 5. Siitä nähdään, että valtaosassa tehtäviä suoritusaika jää alle 20 yksikön, mutta datassa esiintyy tehtäviä, joiden suoritusaika on moninkertainen. Kuvasta nähdään myös, ettei suoritusaika laske monotonisesti vaan n. 30-150 yksikköä kestäviä tehtäviä on myös kohtuullisen paljon. Sovelluskohdetta tuntemalla voidaan histogrammista päätellä, millaisiin tapauksiin datan analyysissä kannattaa paneutua eli millaisten tapausten havaitsemisesta ja optimoinnista voisi olla eniten hyötyä, kun halutaan järjestelmää kehittää ja tarjota asiakkaalle työkaluja järjestelmän toiminnan tehostamiseksi asiakkaan omassa ympäristössä. Esimerkiksi Kuvan 5 tapauksessa voidaan pyrkiä mallinnuksen avulla selvittämään syitä sille, miksi suoritusajan esiintymistiheys ei laske monotonisesti 30 aikayksikön ympärillä tai vaihtoehtoisesti voidaan keskittyä suhteellisen pienemmän määrän kestoltaan ylipitkien tehtävien karsimiseen. Valintaan vaikuttaa se, kumpi on tärkeämpää, vasteaika tai keskimääräinen tehokkuus.



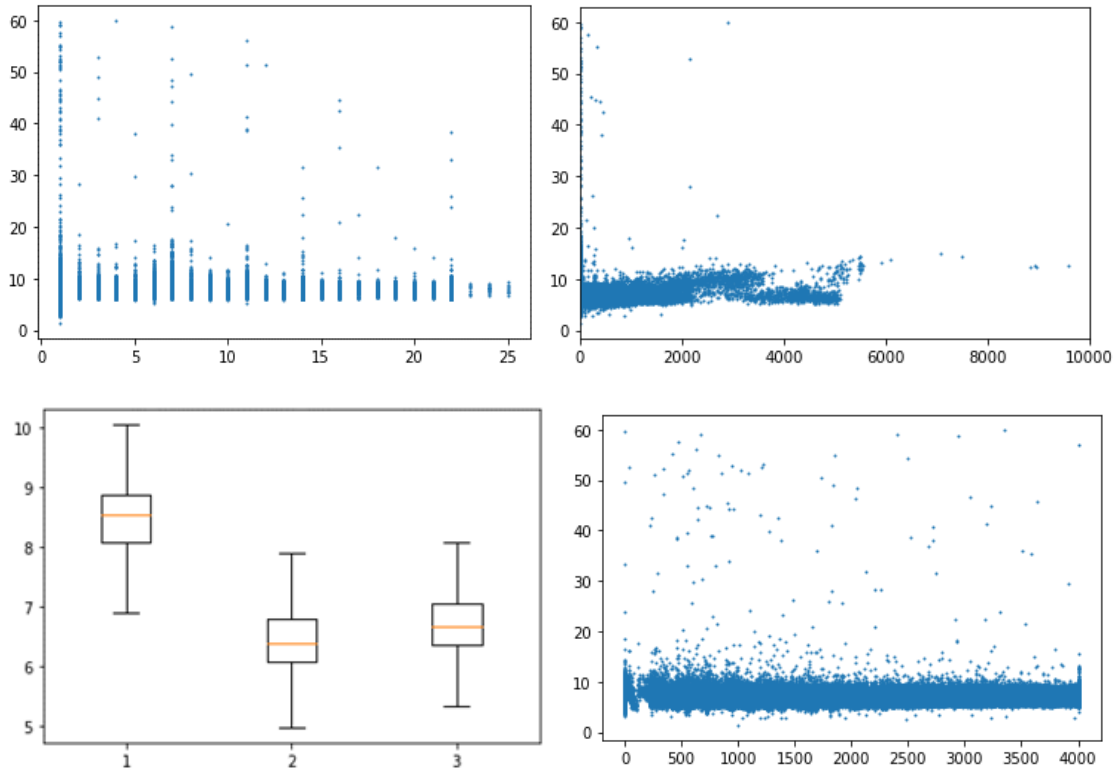
Kuva 5. *Histogrammikuva suoritusajoista*

Histogrammin pohjalta voidaan laskea useita suoritusta kuvaavia parametreja, kuten esim. keskimääräinen suoritusaika, mediaani suoritusaika tai aika, josta nopeammin suoriutuu 90 % tehtävistä (90. persentiili). Tällaiset parametrit ovat erityisen hyödyllisiä, kun verrataan suoritusaikaa esim. ennen ja jälkeen tehostamistoimenpiteen. Myös histogrammin muoto kertoo tarkasteltavasta muuttujasta. Kuvassa 5 oleva histogrammi seuraa kohtuullisen hyvin ns. *log-normal*-jakautumaa (lukuun ottamatta n. 50 yksikön kohdalla olevaa kyhmyä). Tilastotieteestä tiedetään, että tällainen jakauma on luonnollisilla prosesseilla, jotka ovat useamman riippumattoman lähdemuuttujan tulo. Tällainen pohdinta voi kertoa kohdemuuttujan syntymekanismista.

Esimerkkejä muuttujien välisistä korrelaatioista on esitetty Kuvassa 6. Pystyakselilla on kohdemuuttuja ja vaaka-akselilla erityyppisiä selittäviä muuttujia. Vasemmalla ylhäällä oleva kuva esittää kohdemuuttujan korrelaatiota diskreetin lähdemuuttujan kanssa (ts. lähdemuuttuja voi saada vain tiettyjä

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

arvoja; esimerkiksi käsiteltävän laatikkopinon korkeus laatikkojen lukumäärän mukaan). Jos lähdemuuttuja on binääri tai voi saada vain muutaman arvon, voidaan riippuvuus kuvata *boxplot*-kuvaajan muodossa (vasemmalla alhaalla). Kuvassa oranssi keskiviiva kuvaa kohdemuuttujan arvojen mediaania lähdemuuttujan kunkin arvon tapauksessa ja 'viikset' ilmaisevat vaihteluväliä (yleisimmin kvartiileja). Kuvassa 6 esitetyssä tapauksessa kohdemuuttujan arvo selvästi riippuu siitä, onko lähdemuuttujan arvo 1 tai enemmän. Kuvassa 6 oikealla on kuvaajat, jossa lähdemuuttuja on jatkuva (voi saada millaisia arvoja tahansa). Ylemmästä kuvasta nähdään, että muuttujien välinen korrelaatio käyttäytyy eri tavalla lähdemuuttujan eri arvoalueilla, vaikka muuttujien välillä ei olekaan merkittävää korrelaatiota. Oikealla alhaalla olevassa kuvassa kohdemuuttujan arvo on pääasiallisesti sama riippumatta lähdemuuttujan arvosta, mutta koko lähdemuuttujan arvoalueella esiintyy yksittäisiä tapauksia, jolloin kohdemuuttujan (tehtävän suoritusaika) arvot ovat poikkeuksellisen isoja. Tällaisessa tarkastelussa on huomioitava, että yksittäisten lähdemuuttujien korrelaatio (tai korreloimattomuus) kohdemuuttujan kanssa ei kerro vielä, kuinka hyvin kohdemuuttuja on selitettävissä lähdemuuttujien kanssa vaan monimutkaisemmat mallit pystyvät ottamaan huomioon myös kohdemuuttujien epälineaarisia yhteisvaikutuksia (katso mallien kuvaus kappaleessa 6).



Kuva 6. *Erityyppisten muuttujien ja kohdemuuttujan välisen kytköksen visualisointia.*

Myös kahden lähdemuuttujan välisiä korrelaatioita kannattaa tarkastella. Jotkut mallit edellyttävät, että niiden käyttämät lähdemuuttujat olisivat riippumattomia (samalla myös korreloimattomia). Usein tällaisia malleja kuitenkin käytetään, vaikka riippumattomuusehto ei pitääkään täysin paikkansa. Toinen syy tarkastella lähdemuuttujien välisiä korrelaatioita on selvittää, millaiset muuttujat ovat mahdollisesti ylimääräisiä, jos esim. mallinnuksessa on tarvetta karsia syötteitä.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

5. Data-analyysityökalut

Data-analytiikkaa voidaan tehdä melkein kaikissa ohjelmistoympäristöissä, mutta tässä kappaleessa rajoitutaan konkreettisesti jo yrityksellä mahdollisesti olemassa oleviin tai yleisesti saataviin tuotteisiin.

Datan taulukkoesimerkit kuvissa 2-3 on tehty Microsoft Excelillä. Excelissä on hyvin monipuolinen valikoima erilaisia datankäsittely, -puhdistus ja visualisointimenetelmiä, ja kokenut Excelin käyttäjä osaa tehdä nopeasti datasta erilaisia visualisointeja. Tämä on yleensä hyvä alkuvaiheen suunnitteluun ja keskustelun aloittamiseksi datasta ja prosesseista, jota data kuvaa. Lisäksi jos lineaarinen mallinnus on riittävä, Excelissä on valmis työkalu tähänkin tarkoitukseen. Pieni esimerkki lineaarisesta mallinnuksesta Excelillä on viitteessä [2]. Excel on käytössä monissa yrityksissä Microsoft Office paketin myötä.

Jos datan mallinnukseen halutaan esimerkiksi epälineaarisuutta, joustavuutta tai laskentaa vaativia visualisointimahdollisuuksia, eräs vaihtoehto on ilmaiseksi saatava Python ohjelmointikieli ja sille saatavissa olevat erilaiset ilmaiset Python ohjelmointiympäristöt. Kaksi yleisintä Python-ohjelmistoympäristöä ovat Python [3] ja Anaconda [4]. Python ohjelmistoympäristö on pelkkä alusta, joka vaatii asiantuntevan koodaajan data-analytiikkaan tarvittavien ohjelmien tekoon. Python-alustoille on olemassa laaja valikoima valmiiksi tehtyjä kirjastoja, mukaan lukien epälineaariset mallit, SHAP- sekä herkkyystarkastelukirjastot (katso kappale 6). Pythonin kaltaisiin ilmaisohjelmistoihin voidaan laskea myös R- [5] sekä Octave-paketit [6].

Yleisempiä datan analyysiohjelmistoja, joissa on graafinen käyttöliittymä, ovat Tableau, Power BI ja Grafana. Näissä ohjelmistoissa on valmiina yleisimmät datan puhdistus- ja aggregointityökalut erilaisiin valikkorakenteisiin, ja joissakin on valmiina myös yhteenveto- ja mallinnustyökaluja. Lista erilaisista ohjelmistoista löytyy viitteestä [7].

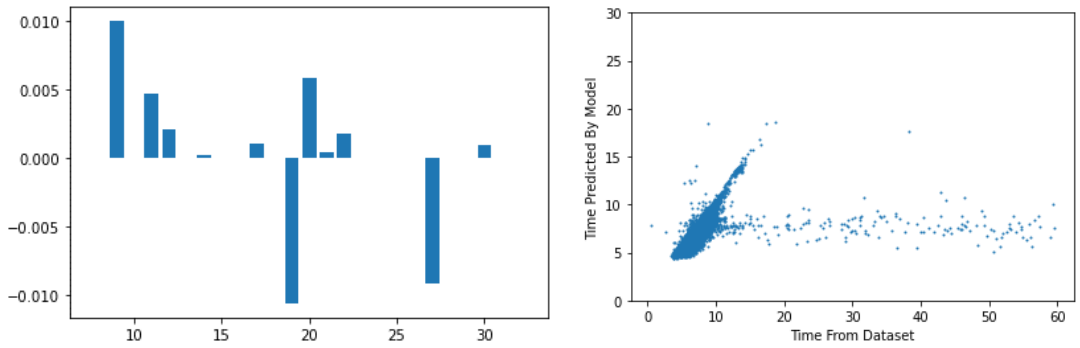
6. Pilotin toteutuksessa käytetyt mallinnusmenetelmät

Datan mallinnusta voidaan käyttää useisiin eri tarkoituksiin optimoinnissa. Perusideana on selittää tietty kohdemuuttuja (ulostulo, y) joukolla muita muuttujia (syötteet, $X = [x_1, x_2, \dots, x_N]$) mallin avulla joka rakenteensa mukaisesti toteuttaa parhaan sovituksen syötteiden ja ulostulon välille. Mallinnusta käytetään yleensä kahteen erityyppiseen tarkoitukseen. Regressiomallinnus on käytössä silloin kun etsitään suoraa yhteyttä mitattujen kanavien (= muuttujien) ja esimerkiksi suoritusajan suhteen. Regression tapauksessa kohdemuuttuja sekä mallin ulostulo ovat arvoasteikoltaan jatkuvia (ts. voivat saada millaisia arvoja tahansa). Toisessa tapauksessa data luokitellaan kategorioihin (esimerkiksi luokitellaan tuotantolinjalta tuleva tuote eri laatukategorioihin, jolloin mallin tarkoituksen on selvittää, millaiset tekijät vaikuttavat tuotteen laatuun sekä arvioida laatukategoriaa joukon muuttujien perusteella). Myös jatkuvan kohdemuuttujan tapauksessa voidaan käyttää luokittelijaa, jos esim. (jatkuva) suoritusaika jaetaan kynnysarvo(je)n perusteella eri arvoalueisiin (esimerkiksi alle/yli 10 sekuntia). Mallilla voidaan myös ennustaa suoritusaikojen tulevaa käyttäytymistä, sekä hakea lopputulokseen eniten vaikuttavia parametreja. Mallin valinta riippuu mallinnettavasta ilmiöstä (esim. muuttujien välisten kytkösten epälineaarisuus). On suositeltavaa aloittaa pienellä yksinkertaisella mallilla ja siirtyä tarvittaessa laajempiin ja monimutkaisempiin malleihin. Seuraavassa tarkastellaan joitakin mallinnusvaihtoehtoja, joista useita käytettiin myös pilotin yhteydessä.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

6.1 Lineaarinen regressio

Lineaarinen (regressio-) mallinnus on menetelmä, jossa kohdemuuttuja ilmaistaan lähdemuuttujien painotettuna summana. Lineaarinen regressiomalli oppii datasta, millaiset lähdemuuttujien kertoimet tuottavat parhaan ulostulon eli millaisilla kertoimilla mallin ulostulon ja kohdemuuttujan todellisen arvon välinen virhe on pienin. Yleisin virhekriteeri on neliösumma. Malli on hyvin yleinen, helposti ymmärrettävä ja laskennallisesti kevyt. Kuvassa 7 on esitetty pilottikohteen datan pohjalta optimoidun lineaarisen regressiomallin kertoimet (vasemmalla) sekä mallin ulostulon ja kohdemuuttujan todellisen arvon välinen korrelaatio.



Kuva 7. *Regressiomallin kertoimet ja mallinnustulos*

Esimerkkitapauksessa malli toimii kohtuullisen hyvin suurimmalle osalle datasta. Malli ei kuitenkaan pysty mallintamaan datassa esiintyviä poikkeuksellisen korkeita kohdemuuttujan arvoja. Vasemmanpuoleisesta kuvasta nähdään, että eniten merkitseviä muuttujia ovat muuttujat 9, 19 ja 27, koska niillä on mallissa absoluuttiarvoltaan suurin kerroin.

Yllä esitetty mallin on lineaarinen eli se ilmaisee kohdemuuttujan lähdemuuttujien painotettuna summana. Joskus kohdemuuttuja käyttäytyy epälineaarisesti, jolloin lineaarinen regressiomalli ei välttämättä ennusta kohdemuuttujan käyttäytymistä riittävän tarkasti.

6.2 Epälineaarinen mallinnus

Epälineaarinen mallinnus kattaa kokoelman erilaisia epälineaarisia ominaisuuksia sisältäviä malleja sekä regressio- että luokitteluongelman ratkaisemiseksi. Lineaarinen regressiomallikin saadaan epälineaariseksi kohdistamalla lähdemuuttujille jokin epälineaarinen operaatio (esimerkiksi logaritmi tai toinen potenssi) ennen malliin syöttämistä.

Epälineaarisiin mallinnusmenetelmiin kuuluvat esimerkiksi:

- päätöspuut (*decision trees*; esim. random forest, XGBoost)
- tukivektorikoneet (*support vector machines*)
- lähimmän naapurin menetelmät (*k-nearest neighbor*)
- neuroverkot (esim. multilayer perceptron, convolutional neural networks jne)

Kun data on esikäsitelty ja tarkasteltu sekä järjestetty taulukkomuotoon, siihen voi helposti sovittaa erilaisia malleja esim. käyttämällä Pythonin scikit-learn -kirjastoa. Mallin valinta ei ole itsestäänselvyys, sillä mallien suorituskyky riippuu datan tyypistä (binääri, kokonaislukudata, kategorinen data ja niin edelleen) sekä mallinnettavasta ilmiöstä.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

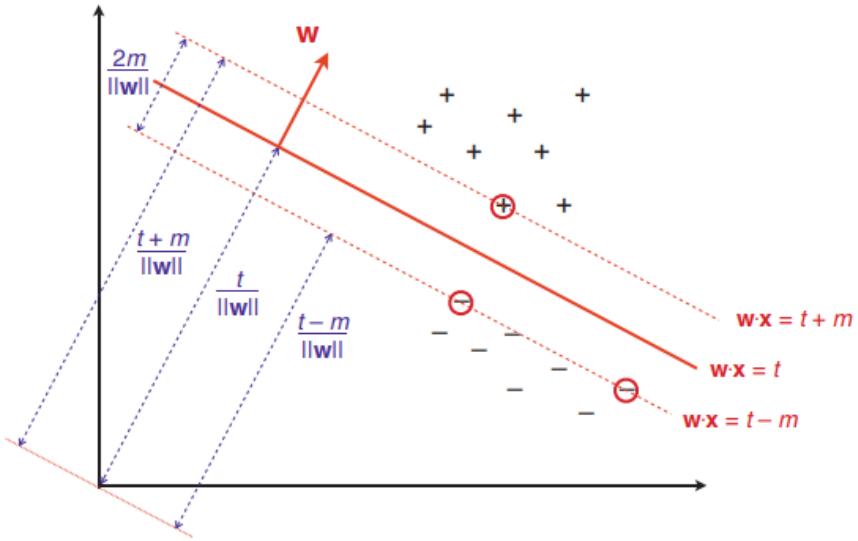
6.2.1 Päätöspuut

Päätöspuu on hierarkkinen malli, jonka pohjana on puumainen rakenne ja jokaisessa solmussa on päätösehto, joka toteutuu tai ei toteudu. Päätösehto määrittää suunnan seuraavalle haaralle (mikäli se on tarpeen). Kun edetään lehtiin päin, päätösehdot muuttuvat yhä yksityiskohtaisemmiksi, kunnes lehtitasolla lopullinen päätös on saavutettu.

Päätöspuita optimoidaan nykyään tietokoneella, jolle annetaan lähtödata ja haluttu ulostulo. Tietokone hakee sopivat ehdot jokaiseen haaraan ja optimoi haarojen ja lehtien määrän. Mikäli päätöspuulta halutaan tilastollista lopputulosta, sisäänmeno- ja ulostulodata voidaan jakaa pareittain satunnaisesti osiin, ja syöttää nämä osadatat useammalle eri päätöspuulle. Eri puut luokittelevat datan hieman eri säännöillä, ja mallin opetusvaiheessa löydetään testidatan perusteella paras kombinaatio eri puiden ulostuloista. Mallia, joka koostuu rinnakkaisista alamalleista, kutsuntaan *ensemble*-malliksi. Päätöspuita hyödyntävä *ensemble*-malli on esimerkiksi satunnaismetsämalli *(random forest*). Pilotissa käytettiin kehittynyttä XGBoost- (*eXtreme Gradient Boosting*) päätöspuurakennetta, joka käsittää algoritmeja päätöspuiden optimointiin.

6.2.2 Tukivektorikoneet

Tukivektorikoneisiin perustuvan regression periaate on lähellä lineaarista regressiota, mutta regressiosuoralle annetaan enemmän toleranssia muodostamalla suoran ympärille toleranssivyöhykkeet. Vyöhykkeet edustavat hyväksyttävää virhettä (ε) ja vyöhykkeiden sisälle jääviä opetusdatan näytteitä kutsutaan tukivektoreiksi. Malli pyrkii minimoimaan yksittäisten pisteiden etäisyyden tukivektoreihin (ξ_i), samalla mallintaen koko pistejoukon mahdollisimman hyvän sovituksen suoralle. Esimerkki tukivektoreista on esitetty Kuvassa 8.



Kuva 8. *Tukivektoriregressio. Kuva on viitteestä [11].*

Suoran hakeminen ei ole pienimmän neliösumman sovitus, vaan iteratiivinen prosessi sopivien kertoimien, hyväksyttävän virheen ja toleranssin löytämiseksi. Kuten lineaarisessa regressiossa, tukivektorikoneilla voidaan ratkaista monimuuttujaongelmia. Tukivektorikoneiden hyvänä puolena voidaan pitää parempaa toleranssia kaukaisten, niin sanottujen *outliers*-pisteiden suhteen verrattuna lineaariseen regressiomenetelmään.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

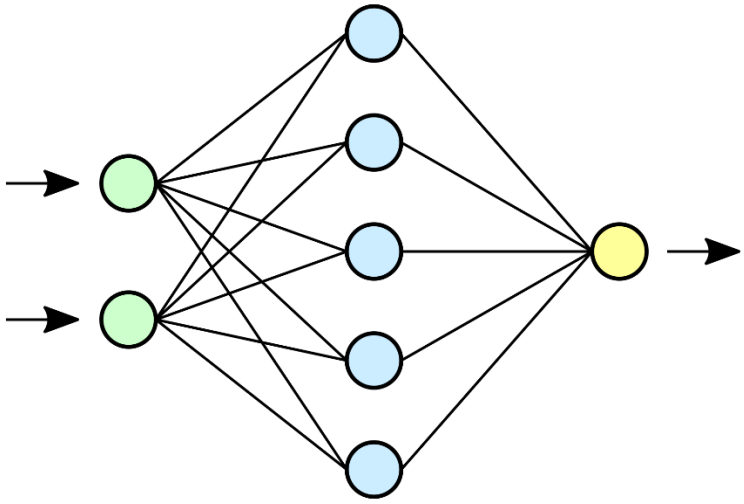
Perusmuodossaan tukivektorikone luokittelee datan kahteen luokkaan (ns. binääriluokittelija) lineaarisesti (ts. luokat erotetaan toisistaan suoralla). Käyttämällä kerneleitä luokkien välisistä rajaviivoista saadaan monimutkaisempia, jolloin pystytään luokittelemaan myös sellainen data, jota kyseisessä parametriavaruudessa ei voida lineaarisesti jakaa luokkiin. Useampitasoisella tukivektorikoneella voidaan jakaa data useampaan kuin kahteen luokkaan. Pilotissa käytettiin tukivektorikoneen *Support Vector Regressor* -algoritmia, joka mahdollistaa käyttäjän määritellä haluttu virhetoleranssi.

6.2.3 K-lähimmän naapurin menetelmä

K lähimmän naapurin regressio- tai luokittelualgoritmin parametrina on naapureiden määrä. Algoritmi hyödyntää opetusdataa, jonka tapauksessa vaste (=kohdemuuttujan arvo) syötteille on tiedossa. Kohdemuuttujan arvo tuntemattomalle syötteelle saadaan löytämällä sille *k* lähintä syötevektoria opetusdatasta ja keskiarvottamalla niitä vastaavat kohdemuuttujan arvot. Algoritmissa voidaan vaihdella naapureiden määrää ja etäisyyden laskentaa (esimerkiksi euklidinen tai Manhattan-etäisyys) parhaan lopputuloksen saamiseksi.

6.2.4 Neuroverkot

Neuroverkot ovat tiedon käsittelyn malleja, jotka ovat saaneet inspiraationsa aivojen rakenteesta. Neuroverkko on topologia, joka koostuu laskenta-alkioista, joita kutsutaan neuroneiksi tai solmuiksi. Neuroverkolle syötetään sisääntulotiedot ja yksittäinen solmu laskee niistä painotetun summan. Solmun ulostulo normalisoidaan käyttämällä epälineaarista aktivointifunktiota. Solmuja voi olla useita rinnakkain ja monessa kerroksessa, jolloin edellisen kerroksen ulostulot ovat seuraavan kerroksen syötteinä. Esimerkki yksinkertaisesta neuroverkosta on Kuvassa 9.



***Kuva 9**. Neuroverkko, kuva otettu viitteestä [10].*

Neuroverkot ovat ohjatun ja ohjaamattoman oppimisen laskentamalleja. Ohjatussa oppimisessa hyödynnetään opetusdataa, jonka tapauksessa haluttu ulostulo tiedetään. Neuroverkkomalli säätää sisäisiä parametrejään ’oppimalla’ opetusdatasta ja minimoimalla verkon ulostulon ja oikean vasteen välistä virhettä. Ohjaamattomassa oppimisessa ulostuloa ei tiedetä, ja verkko pyrkii ryhmittelemään dataa jonkin minimilaskusäännön mukaisesti. Neuroverkkojen etuna on niiden joustavuus ja nopea laskenta rinnakkaistamalla tai mahdollisesti grafiikkakiihdyttimen avulla.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

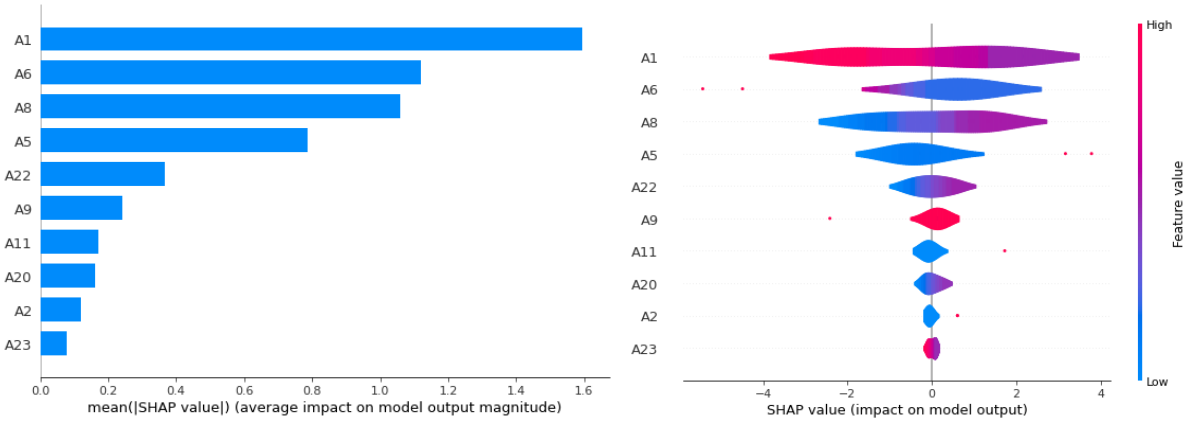
Yllä mainituista mallinnusmenetelmistä löytyy runsaasti kirjallisuutta eikä niiden yksityiskohtaisempi kuvaus mahdu tähän raporttiin. Hyvä yhteenveto erilaisista mallinnusmenetelmistä on viitteessä [10]. Pilotissa testattiin useiden eri mallinnusmenetelmien soveltuvuutta pilottikohteen datan mallintamiseen. Näitä olivat muun muassa lineaariset regressiomallit variantteineen, päätöspuut, K-lähimmän naapurin mallinnus, tukivektorikoneet sekä neuroverkot. Lähes kaikki mallit sisältävät niin sanottuja säätö- eli hyperparametrejä ja mallien optimointiin on hyvä kiinnittää huomiota. Pilottikohteessa ajanpuutteen vuoksi pystyttiin optimoimaan vain muutamaa mallia.

6.3 Mallin validointi

Mallin validoinnissa mallin suorituskykyä tarkastellaan kriittisesti kuvien ja laskettujen mittareiden avulla. Mallien sorituskykyä arvioitaessa voidaan käyttää *coefficient of determination* eli R^2 -mittaria. Se kertoo mallin hyvyyden reaalilukuna -∞…0…1, jolloin numero 1 vastaa täydellistä mallinnusta lähtöarvojen ja kohdemuuttujan välillä. Vastaavasti nollatulos tarkoittaa, että mallin ennuste vastaa kohdemuuttujan keskiarvoa (toisin sanoen, mallista ei ole ei hyötyä eikä haittaa). R^2-arvo voi myös olla negatiivinen, jolloin mallin tuottama ennuste tai estimaatti on erisuuntainen kuin mallinnettavan kohdemuuttujan oikea arvo.

Muita vastaavia mittareita ovat *coefficient of correlation*, joka antaa mallin hyvyysalueeksi -1…1. Luku -1 tarkoittaa täydellistä negatiivista korrelaatiota mallin lähtöarvojen ja ulostulon välillä, ja luku 1 tarkoittaa täydellistä positiivista korrelaatiota. Luku 0 vastaa mallin kyvyttömyyttä löytää korrelaatiota lähdön ja ulostulon parametrien välille.

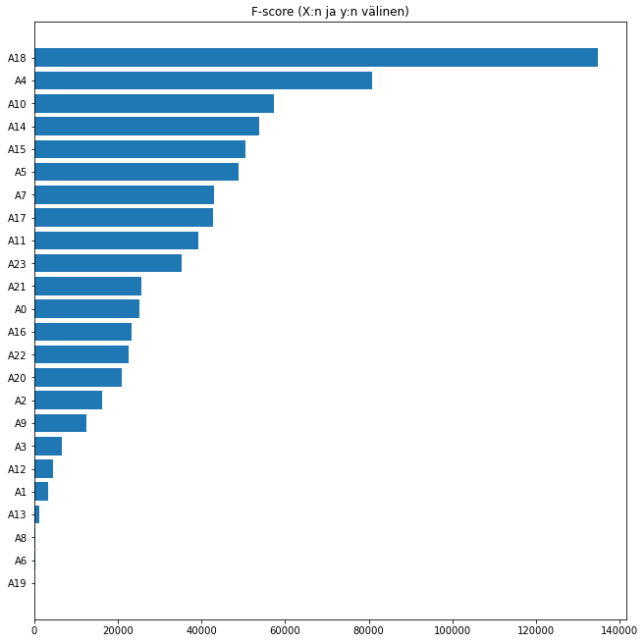
Mallin syötteiden vaikutusta lopputulokseen voidaan arvioida myös esimerkiksi Shapleyn [11] menetelmän avulla. Syötteiden yksittäinen tai keskinäinen vaikutus lopputulokseen lasketaan ja visualisoidaan useilla eri tekniikoilla, jotka osittain liittyvät myös mallin valintaan [12]. Kuvassa 10 on kaksi esimerkkiä parametrien vaikutuksien arvioinnista mallin ulostuloon Shapleyn menetelmällä.



***Kuva 10**. SHAPley visualisointeja pilottidatalle. Vasemmalla on keskimääräinen parametrien vaikutus ulostuloon ja oikealla merkittävimmät parametrit ovat merkitty punaisella, vähän merkitsevät sinisellä, ja parametrien jakauma paksummalla osalla parametrilinjaa.*

Syöttödatan vaikutusta (joko mallin kautta tai ilman) lopputulokseen voidaan myös tarkastella ***F-scoren*** (joskus myös F1-score) avulla. Tässä menetelmässä datan jokaista saraketta verrataan tilastollisena histogrammina ulostulon histogrammiin *chi-square* (χ^2) laskentamenetelmän avulla. Näin saadut kertoimet järjestetään suuruusjärjestykseen ja voidaan esittää esimerkiksi Kuvan 11 tavoin.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.



***Kuva 11**. F-score analyysin tulos*

Mallien parametrien käyttäytymistä voidaan tarkastella myös herkkyystarkastelumenetelmien (*sensitivity index*, esimerkiksi *Sobol sensitivity index*) avulla [13]. Niissä tarkastelun kohteena on yksittäisen kanavan vaihtelu ja kuinka se vaikuttaa lopputuloksen (esimerkiksi tehtävän suoritusaika) vaihteluun. Ensimmäisen asteen (*first-order*) herkkyystarkastelu osoittaa mallin ulostulon herkkyyttä parametrien arvojen vaihteluun, kun tarkastellaan aina yhtä parametriä kerrallaan. Kokonaistarkastelu (*total-order sensitivity analysis*) sen sijaan pyrkii ottamaan huomioon parametrien keskinäiset kytkökset. Jos molemmat antavat lähes samanlaiset tulokset, voidaan päätellä, että parametrit ovat keskenään riippumattomia, ainakin kyseisen mallin näkökulmasta.

Mallin tulokset tulee olla ymmärrettäviä ja loogisesti pääteltävissä myös yrityksen avainhenkilöiden toimesta. Mikäli malli ei selitä dataa kunnolla tai tulokset herättävät epäilyjä, on mallia iteroitava tai kokeiltava jotakin muuta mallinnustapaa. Mallin antaessa intuitiivisia, odotuksia vastaavia regressio- tai luokittelutuloksia voidaan mallinnus katsoa hyväksytyksi.

Pilotissa tarkasteltiin useiden eri mallien hyvyyttä sekä niiden herkkyyttä syötteille. Tarkoituksena oli selvittää, millaiset robottijärjestelmän eri tehtäviä kuvaavat ominaisuudet vaikuttavat eniten tehtävän suoritusaikaan. Todettiin, että eri mallinnusmenetelmillä mallin syötteiden merkittävyys vaihteli, joskin tietyt syötteet sijoittuivat yleensä aina tärkeimpien joukkoon.

7. Data-analytiikkapilotin konsepti ja johtopäätökset

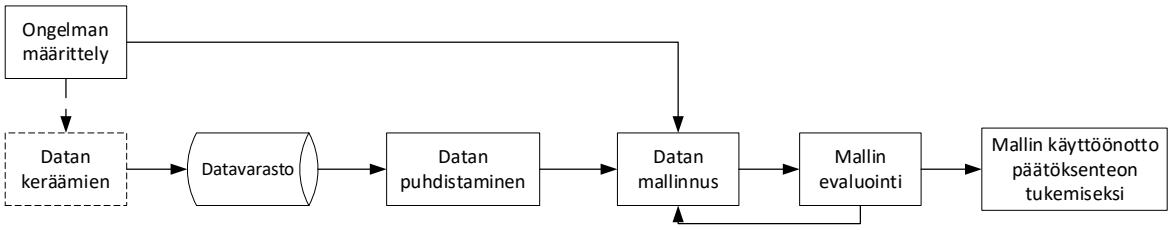
Johtopäätöksissä tarkastellaan ensin data-analytiikkaprojektin etenemisprosessia, jonka jälkeen tuodaan esiin muutamia tärkeimpiä pilotista opittuja seikkoja.

7.1 Data-analytiikkapilotin konseptin kuvaus

Pilotissa toteutettu robottijärjestelmän datalähtöinen analyysi on tyypillinen data-analytiikkaprojekti, jossa datan pohjalta selvitetään jonkun laitteen, tuotantolinjan tai ilmiön toimintaa ja siihen vaikuttavia tekijöitä. Projekti koostuu sarjasta loogisessa järjestyksessä suoritettavia tehtäviä (Kuva 12). Ensin

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

määritellään ongelma, mitä halutaan tutkia tai optimoida järjestelmässä. Tämän jälkeen katsotaan, onko läheisesti ongelmaan liittyviä muuttujia mitattu ja tallennettu datavarastoon. Data puhdistetaan tarvittaessa ja tämän jälkeen haetaan lähestymistapaa mallinnukseen, esimerkiksi lineaarista tai epälineaarista mallinnusta. Mallinnusta arvioitaessa datan visualisointi on välttämätön vaihe, sillä se voi paljastaa järjestelmästä seikkoja, jotka vaativat toimenpiteitä jo ennen mallinnusta. Mallinnus voi olla iteratiivinen prosessi, ja mallin evaluointi sekä mallinnustuloksien tarkastelu on tehtävä yhdessä yrityksen asiantuntijoiden kanssa. Kun malli on todettu toiminnallisesti hyväksi, se voidaan ottaa tuotantokäyttöön ja tehdä sen perusteella päätöksiä linjaston toiminnan optimoinnista. Linjasto tarkoittaa tässä yhteydessä kappaleen tai objektin kulkua prosessin läpi, joka sisältää robottien avulla tehtäviä toimenpiteitä kappaleelle tai objektille. Kuvassa 12 on esitetty analytiikka- ja optimointiprosessin kuvaus.



***Kuva 12**. Konseptin prosessikuvaus*

7.1.1 Ongelman määrittely

Ongelman määrittely on keskeisessä osassa projektia. Siihen vaikuttavat datan kanavien määrä, datan keräysaika sekä kuinka paljon projektissa on aikaa tutkia kyseistä ongelmaa. Nämä yhdessä vaikuttavat keskeisesti projektin tavoitteisiin saada haluttu lopputulos. Tässä pilotissa ongelmaksi oli asetettu automaatiojärjestelmän kapasiteetin seuranta ja tavoitteena oli data-analytiikan avulla löytää sellaisia tekijöitä, jotka vaikuttavat järjestelmän kapasiteettiin.

7.1.2 Datan kerääminen

Datan kerääminen on yleensä automatisoitu prosessi, jossa sensoreilta saatu data kirjoitetaan lokijärjestelmään. Joissakin tapauksissa osa tiedoista syötetään tiedostojärjestelmään käsin, mikä voi johtaa unohtamisiin ja väärien tietojen syöttämisiin. Yleensä tietojen käsin syötöstä tulisi pyrkiä eroon edellä mainittujen virheiden poistamiseksi.

7.1.3 Datavarasto

Datavarasto on tietokanta, josta voidaan hakea tarvittavat tietokanavat tietylle aikavälille tarkasteltavaksi. Yleensä tietovarasto on varmuuskopioitu ja tietosuojattu varasto yrityksen tiedoille. Datavaraston toteutukseen vaikuttavia tekijöitä ovat mm. vaadittavat tietosuojakäytännöt, datan määrä, datan kirjoitus- ja lukuoperaatioiden muoto (esim. luetaanko/kirjoitetaanko dataa järjestyksessä tai tarvitaanko satunnainen pääsy dataan), reaaliaikaisuusvaatimukset ym.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

7.1.4 Datan puhdistaminen

Datan puhdistus on välttämätöntä, sillä se paljastaa mahdolliset kirjauksessa tapahtuneet virheet, puutteelliset rajamäärittelyt automaattisissa lokitallenteissa sekä vikaantuneet sensorit järjestelmässä. Virheellisen datan käyttö mallinnuksessa johtaa vaikeuksiin tulosten tulkitsemisessa ja pahimmillaan myös virheellisiin johtopäätöksiin.

7.1.5 Datan mallinnus ja evaluointi

Datan mallinnus on keskeinen osa projektia, koska sen avulla voidaan tuottaa parhaimmillaan digitaalinen malli järjestelmästä ja löytää siten pullonkaulat yrityksen järjestelmästä. Mallia on evaluoitava aika-ajoin, jotta varmistetaan mallin hyvyys ja käyttökelpoisuus suorituskyky- ja kapasiteettiennusteissa.

7.1.6 Mallin käyttöönotto

Kun malli on hyväksytty käyttöön, sillä voidaan tehdä yrityksessä vertailevaa seurantaa mallin tuottamien ennusteiden ja todellisten tulosten välillä. Tulosten avulla voidaan määritellä toimenpiteitä ja keinoja, joilla järjestelmän suorituskykyä voidaan parantaa.

Pilottiprojektissa käytettiin Kuvan 12 lähestymistapaa. 3D Data Economy -hanke oli määritellyt viitekehyksen lähestymistavalle ja kohdeyritykseltä saatiin halutun viitekehyksen sisälle määritelty käytännön ongelmakehys sekä *case*-aineisto pilottiprojektille. Datan kerääminen oli siis jo tehty ja koko ajan käynnissä, ja datan muuttujien valinta analyysiä varten kävi melkein reaaliaikaisesti. Keskeisintä projektissa oli kaikkien osapuolien välinen tiivis viestintä sekä säännölliset palaverit noin viikon välein. Projekti vaatii myös henkilöpanostusta yritykseltä lähinnä asiantuntijatasolta kommenttien ja mielipiteiden muodossa sekä sitoutumista projektiin näiltä osin. Pilottiprojektissa päästiin aloittamaan datan puhdistamisesta ja epärelevanttien kanavien poistosta. Suurin haaste oli löytää mallijoukko ja menetelmät, millä voidaan kuvata parametrien vaikutusta suoritusaikaan. Koska mallit ovat erilaisia (esim. lineaarisia ja epälineaarisia), eri mallien tuottamat tulokset vaihtelevat luonnollisesti jonkun verran. Loppuvaiheessa päädyttiin tarkastelemaan usean mallin tuloksista luotua synteesiä, mistä voidaan verrata eri mallien tuloksia keskenään ja päätellä merkittävimpien parametrien joukko.

7.2 Pilottiprojektin opetukset

Pilottiprojektissa oli huomionarvoista ongelman selkeä määrittely. Selkeä määrittely nopeuttaa liikkeellelähtöä merkittävästi, ja vaikka datasettiä ja muuttujia vaihdetaan, lopputavoite on silti sama koko projektin ajan. Datan puhtaus on myös oleellinen asia projektin nopeassa käynnistymisessä.

Toinen tärkeä seikka pilotin onnistumisen näkökulmasta on tiivis kommunikointi projektin osapuolien kesken. Pilottikohteessa yhteydenpito tapahtui viikkopalaverien avulla, johon kaikki osapuolet olivat kutsuttuna mukaan. Näin tieto ja kehitystyön ongelmat tuotiin heti esille ja kaikille arvioitavaksi, ja ratkaisuehdotuksia pystyttiin pohtimaan yhdessä. Lisäksi päivittäisiä ongelmakohtia ja lisäkysymyksiä jaettiin kaikille osapuolille avoimen (muilta suljetun) keskustelukanavan kautta.

Mallinnuksessa tarvittava laskentakapasiteetti oli peruskannettavan tietokoneen kohdalla riittävä, mutta 32 gigatavun keskusmuisti riitti juuri ja juuri pitämään kaikki 6,5 miljoonaa datariviä työmuistissa.

Kehitystyössä käytettiin kohdeyritykselle tuttua työkalua, joka oli Python-koodausympäristö. Tämä lisäsi osaltaan yhteistyön nopeutta ja joustavuutta, sillä kohdeyrityksestä pystyttiin nopeasti kommentoimaan koodissa olevia puutteita ja antamaan kehitysehdotuksia.

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.

Näistä yhteenvetona voidaan todeta, että selkeä ja avoin kommunikaatio, selkeä ymmärrys menetelmistä ja toimenpiteistä datalle, avoimien järjestelmien käyttö kaikkien osapuolien kesken ovat avainasemassa projektin onnistumisen suhteen.

7.3 Tampereen yliopiston työryhmä

Tässä raportissa kuvattu pilotti toteutettiin Tampereen yliopiston Data-analytiikan ja Optimoinnin (DAO) tutkimusryhmän toimesta ja siihen osallistuivat yliopistonlehtori Jari Turunen (pilotin käytännön toteutus) sekä professori Tarmo Lipping (projektin vastuuhenkilö). Yliopistonlehtori Jari Turusella on pitkäaikainen kokemus data-analytiikkaan liittyvien menetelmien kehittämisestä eri sovellusalueilla. Hän opettaa aikasarjamuotoisen datan analyysimenetelmiä sekä koneoppimista käsitteleviä kursseja ja on osaltaan ohjannut lukuisia diplomitöitä koneoppimiseen ja data-analytiikkaan liittyen. Tarmo Lipping on toiminut signaalinkäsittelyn professorina vuodesta 2004. Hän on ollut vastuullisena johtajana lukuisissa signaalinkäsittelyyn ja data-analytiikkaan liittyvissä projekteissa sekä ohjannut yli 50 diplomityötä ja 7 väitöskirjaa.

Viitteet

[1] https://media.cimcorp.com/web/1bad4b2bf43583/press-images/?viewType=grid , 24.4.2023

[2] https://www.ablebits.com/office-addins-blog/linear-regression-analysis-excel/ , 19.5.2023

[3] https://www.python.org/ , 19.5.2023

[4] https://anaconda.org/ , 19.5.2023

[5] https://cran.r-project.org/ ,19.5.2023

[6] https://www.octave.org/ , 19.5.2023

[7] https://www.getapp.com/business-intelligence-analytics-software/data-analysis/p/free/ , 19.5.2023

[8] Korpua, A. 2019. Gradienttitehostetut päätöspuut. Pro Gradu tutkielma, Turun yliopisto

[9] https://fi.wikipedia.org/wiki/Neuroverkot , 6.6.2023

[10] Flach, P. 2012. Machine Learning: The Art and Science of Algorithms That Make Sense of Data, Cambridge University Press

[11] https://en.wikipedia.org/wiki/Shapley_value , 27.4.2023

[12] https://shap.readthedocs.io/en/latest/index.html , 27.4.2023

[13] https://en.wikipedia.org/wiki/Sensitivity_analysis , 27.4.2023

Hanke rahoitetaan REACT-EU-välineen määrärahoista osana Euroopan unionin COVID-19-pandemian johdosta toteuttamia toimia.