Mikko Koponen & Virpi Hotti

Tullidata – Excel-työkalu
ja suoritettava asiakirja (Colab)

KARELIA-AMMATTIKORKEAKOULU

Tullidata – Excel-työkalu ja suoritettava asiakirja (Colab)

Tekijät: Mikko Koponen & Virpi Hotti (Itä-Suomen yliopisto, tietojenkäsittelytieteen laitos)
Sivuntaitto: Kaisa Varis
Kansikuva: STANLEY NGUMA from Pexels / Unsplash
Kustantaja: Karelia-ammattikorkeakoulu & KoDa – datan kokonaisvaltainen hallinnointi ja hyödyntäminen -hanke, 2020

ISBN:978-952-275-302-1

Karelia AMMATTIKORKEAKOULU

UNIVERSITY OF EASTERN FINLAND

BUSINESS JOENSUU | JOE WITH

Euroopan unioni
Euroopan aluekehitysrahasto
Euroopan sosiaalirahasto

Vipuvoimaa EU:lta 2014–2020

Elinkeino-, liikenne- ja ympäristökeskus

Sisällys

Esipuhe

Yrityksille on hyödyksi, jos he pystyvät ennakoimaan ihmisten ja tavaroiden liikkumista yrityksen toiminta-alueella. Tässä julkaisussa on kuvattu **KoDa – Kokonaisvaltainen datan hallinnointi ja hyödyntäminen (ESR 2017-2020)**[1] -hankkeen yhdessä pilotissa käytetyt työkalut ja niiden käyttö. Pilotin tavoitteena oli soveltaa tiedon keräys- ja analysointimenetelmiä suurten ihmis- ja tavaramäärien liikkumisen ennakointiin Niiralan rajanylityspaikalla. Excel-työkalun avulla tuonti- ja vientimääriä pystyy tutkimaan valituilla hakuehdoilla. Työkalu hyödyntää ilmaista, verkosta saatavaa dataa niin, että käytössä on aina viimeisin saatavilla oleva data. KoDa-hankkeen verkkoisivuilta[2] ladattavissa oleva työkalu on vapaasti kaikkien aiheesta kiinnostuneiden hyödynnettävissä ja muokattavissa. Sen ovat kehittäneet ja tämän julkaisun kirjoittaneet Itä-Suomen yliopiston erityisasiantuntija Mikko Koponen ja lehtori Virpi Hotti.

Säde Lind, projektipäällikkö
Karelia-ammattikorkeakoulu

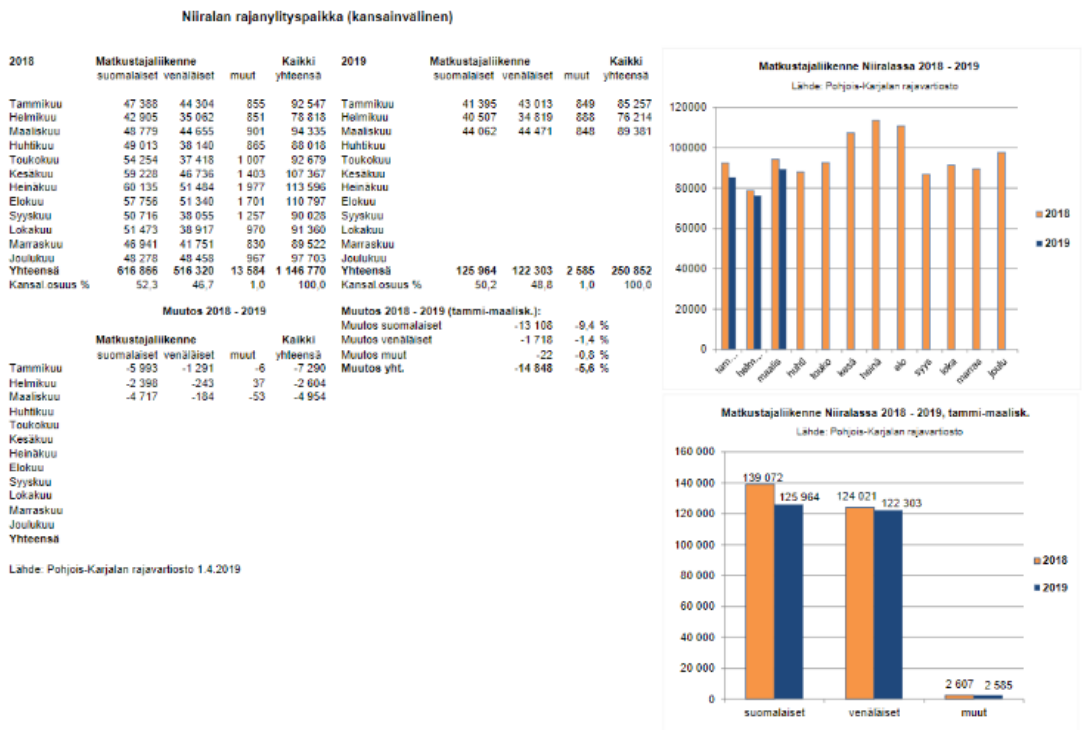
[1] https://www.eura2014.fi/rrtiepa/projekti.php?projektikoodi=S20980
[2] http://www.karelia.fi/koda/tyokalut-2/tulokset/

1 Johdanto

Niiralan pilotissa haluttiin selvittää Niiralan kautta kulkeva tavaraliikenne kokonaisuutenaan, jotta saatuja tietoja voivat hyödyntää tavarantoimittajat ja logistiikkaoperaattorit omassa liiketoiminnassa sekä erinäiset organisaatiot edunvalvonnassa. Jos tarkastellaan kokonaisvaltaisesti tavaraliikennettä, niin silloin tarkastellaan lukumäärä per viikonpäivä ja ajankohta, tuonti ja vienti (mitä viedään ja mitä tuodaan), rautatie ja maantie, mistä tuodaan ja mihin viedään. Esimerkiksi konttikoot (TEUt) voidaan unohtaa, vaikka tavaravirrat yleensä ilmoitetaan TEUna.[3]

Niiralan pilotissa käytetään ainoastaan tullidataa (Luku 2), sillä emme saaneet dataa seuraavilta tahoilta: Global blue, Etaxfree ja TAK. Rajavartiolaitoksen tietovarannosta saadaan rajanylittäneiden määrä ja luokiteltu kansalaisuus (eli onko suomalainen, venäläinen tai muut), mutta data jätettiin ulkopuolelle, sillä sitä ei saada yhdistettyä esimerkiksi avoimeen transitiodataan, joka kertoo tavararyhmät.

Koska Uljas.fi-palvelusta saadaan osa tullidatasta, niin KoDa-hankkeessa toteutettiin Excel-työkalu Mikko Koposen (UEF) toimesta, jonka avulla saadaan tullidatan eri tietokannoista haettua dataa eri tarpeisiin. Excel-työkalua (Luku 3) ovat testanneet Mikko Koponen (UEF), Virpi Hotti (UEF), Mika Lappalainen (Karelia-amk) sekä Säde Lind (Karelia-amk). Suoritettavaa asiakirjaa (Luku 4) ovat testanneet Mikko Koponen ja Virpi Hotti. Niiralan rajanylittäjien määrää ennustettiin automaattista koneoppimista hyödyntämällä (Luku 5).



Niiralan rajanylityspaikka (kansainvälinen)

2018	Matkustajaliikenne suomalaiset	venäläiset	muut	Kaikki yhteensä
Tammikuu	47 388	44 304	855	92 547
Helmikuu	42 905	35 062	851	78 818
Maaliskuu	48 779	44 655	901	94 335
Huhtikuu	49 013	38 140	865	88 018
Toukokuu	54 254	37 418	1 007	92 679
Kesäkuu	59 228	46 736	1 403	107 367
Heinäkuu	60 135	51 484	1 977	113 596
Elokuu	57 756	51 340	1 701	110 797
Syyskuu	50 716	38 055	1 257	90 028
Lokakuu	51 473	38 917	970	91 360
Marraskuu	46 941	41 751	830	89 522
Joulukuu	48 278	48 458	967	97 703
Yhteensä	616 866	516 320	13 584	1 146 770
Kansal.osuus %	52,3	46,7	1,0	100,0

2019	Matkustajaliikenne suomalaiset	venäläiset	muut	Kaikki yhteensä
Tammikuu	41 395	43 013	849	85 257
Helmikuu	40 507	34 819	888	76 214
Maaliskuu	44 062	44 471	848	89 381
Huhtikuu				
Toukokuu				
Kesäkuu				
Heinäkuu				
Elokuu				
Syyskuu				
Lokakuu				
Marraskuu				
Joulukuu				
Yhteensä	125 964	122 303	2 585	250 852
Kansal.osuus %	50,2	48,8	1,0	100,0

Muutos 2018 - 2019

	Matkustajaliikenne suomalaiset	venäläiset	muut	Kaikki yhteensä
Tammikuu	-5 993	-1 291	-6	-7 290
Helmikuu	-2 398	-243	37	-2 604
Maaliskuu	-4 717	-184	-53	-4 954
Huhtikuu				
Toukokuu				
Kesäkuu				
Heinäkuu				
Elokuu				
Syyskuu				
Lokakuu				
Marraskuu				
Joulukuu				
Yhteensä				

Muutos 2018 - 2019 (tammi-maalisk.):

Muutos suomalaiset	-13 108	-9,4 %
Muutos venäläiset	-1 718	-1,4 %
Muutos muut	-22	-0,8 %
Muutos yht.	-14 848	-5,6 %

Lähde: Pohjois-Karjalan rajavartiosto 1.4.2019

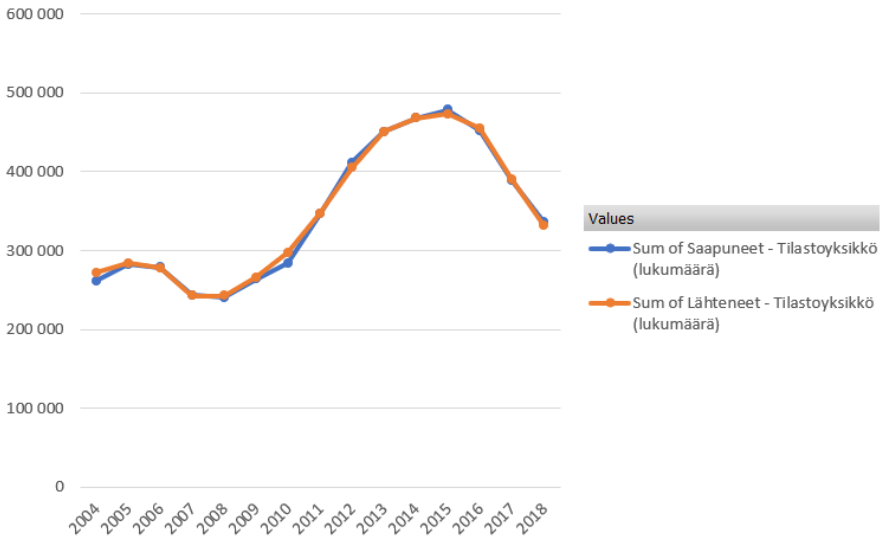
Kuva 1. Esimerkki rajavartiolaitosdatan Niiralan rajanylityspaikan osalta[4].

[3] Mukailtu Birgitta Väisäsen, Keski-Karjalan Kehitysyhtiö Oy KETI, asiantuntijalausunnosta
[4] http://kareliaexpert.fi/2019-04-17/niiralan-rajanylitykset-maaliskuun-2019-lopussa/

2 Tullidata

Uljas.fi-palvelun käyttöön on olemassa ohje[5], jota voidaan hyödyntää, kun haetaan esimerkiksi rajaliikennedataa ja transitiodataa. Esimerkkinä Niiralan rajanylityspaikan rajaliikennedata[6] on saatavilla vuodesta 2004 alkaen kuukausitasolla ja datassa on sekä saapuneiden että lähteneiden liikennevälineiden lukumäärät (Kuva 2). Liikennevälineistä (henkilöauto, kuormattu irtoperävaunu, kuormattu kontti, kuormattu kuorma-auto, linja-auto, tyhjä irtoperävaunu, tyhjä kontti, ja tyhjä kuorma-auto) dataa löytyy kaikille vuosille henkilöautojen, kuormattujen kuorma-autojen ja tyhjien kuorma-autojen osalta. Linja-autojen osalta dataa löytyy vuosille 2004-2009. Tyhjien irtoperävaunujen ja tyhjien konttien osalta ei dataa ole olemassa.

Esimerkkinä Niiralan rajanylityspaikan transitiodataa on saatavilla vuodesta 2011 alkaen esimerkiksi euro- ja kilomäärittäin jaettuna eri tavararyhmille (Kuva 3).



Kuva 2. Esimerkkinä Niiralan rajanylityspaikan kaikki tilastoidut lähteneet ja saapuneet liikennevälineiden (data haettu 7.5.2019)

[5] https://tulli.fi/tilastot/uljas-tilastotietokannan-ohjeita
[6] http://uljas.tulli.fi/uljas/

Tavararyhmä	2011	2012	2013	2014	2015	2016	2017	2018	Grand Total
10 Peruskemikaalit				42 371	1 506 976	677 387	3 497 406	9 620 736	15 344 876
12 Väriaineet			375	4 280	3 396	700	4 314	34 799	47 864
13 Pesuaineet, kosmetiikka- ja toalettivalmisteet		3 307	94	867	1 083				5 351
14 Muu kemiallinen teollisuus	433 242	94 276	727		1 799	32 161	41 097	33 352	636 654
15 Muovit	556 479	238 558	227 153	196 295	94 239	97 931	298 499	616 104	2 325 258
16 Kumi	1 286	127 297	49	1 109	6 892	259	10 657	2 819	150 368
17 Nahka, nahkatuotteet		298				63	124		485
18 Metsäteollisuustuotteet	5 865 176	3 694 919	2 393 875	408 151	136 666	81 670	182 454	131 270	12 894 181
19 Tekstiilit, vaatteet, jalkineet	87 571	438 896	80 619	515 206	301 906	620 137	920 870	1 269 744	4 234 949
2 Kalat, kalatuotteet		10 609	4 125						14 734
20 Kivitavarat, keramiikka, lasi	10 646	29	40	171	1 298		2 949	316	15 449
21 Rauta ja teräs sekä tavarat	119 099	211 794	125 841	2 555	2 510	209 015	338 940	144 021	1 153 775
22 Työkalut	206 865	1 561	23 280	2 957	1 233	1 946	9 424	25 085	272 351
23 Radio-, televisio- ja tietokonelaitteet	11 471	1 332	384	607 391	88 561		789	19 513	729 441
24 Kodinkoneet	304	2 440	410	3 686	619	36 648	151		44 258
25 Muut koneet ja laitteet	16 393 674	13 914 870	9 330 964	6 606 230	3 637 561	580 071	1 261 901	5 029 395	56 754 666
26 Autot	2 424 274	4 612 617	3 042 970	243 222			23 709	55 378	10 402 170
27 Autojen osat, perävaunut, moottori- ja polkupyörät	34 819	38 181	25 888	154 951	1 401	19 861	65 584	108 410	449 095
28 Lääkintäkojeet, mittaus-, tarkkailu- ja optiikkavälinee	79 084	82 951	29 493	4 582	22 583	1 877	13 616	30 053	264 239
29 Huonekalut	984 041	684 792	475 410	688 168	3 268	505	8	6	2 836 198
30 Lelut, pelit, urheiluvälineet	62	413	241	6 200	215	18 594	16 874	529	43 128
31 Muut	20 479	23 605	27 822	10 357	3 061	4 196	13 964	59 441	162 925
32 Erittelemätön	237 421	46 162	57 593	286 672	551 498	134 183	546 805	322 070	2 182 404
4 Tee				5 300	3 089	8 301	16 687	8 550	41 927
6 Sokeri, kaakao, suklaa, makeiset					236	72		100	408
8 Muut elintarvikkeet	242 461	236 048	1 109 922	8 543	15 419	21 293	81 408	215 203	1 930 297
9 Öljytuotteet	32 779								32 779
Grand Total	**27 741 233**	**24 464 955**	**16 957 275**	**9 799 264**	**6 385 509**	**2 546 870**	**7 348 230**	**17 726 894**	**112 970 230**

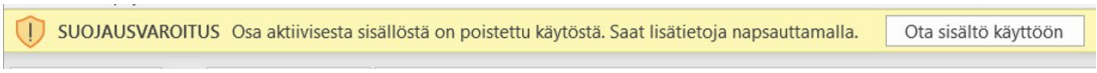
Kuva 3. Esimerkkinä Niiralan rajanylityspaikan kaikki tilastoidut tavararyhmät euroina (data haettu 7.5.2019)

Valittavista tietokannoista löytyy erilaisia Tullin ylläpitämiä tilastoja, jotka ovat tarkemmin kuvattuna taulukossa (Liite 1). Tulli-datasta otetaan vuosilta 2004-2019 Niiralaan saapuneet ja sieltä lähteneet, tuonti- ja vientimäärät, kauppatasetiedot, saapuneet ja lähteneet Norjaan ja Ruotsiin.

3 Excel-työkalu

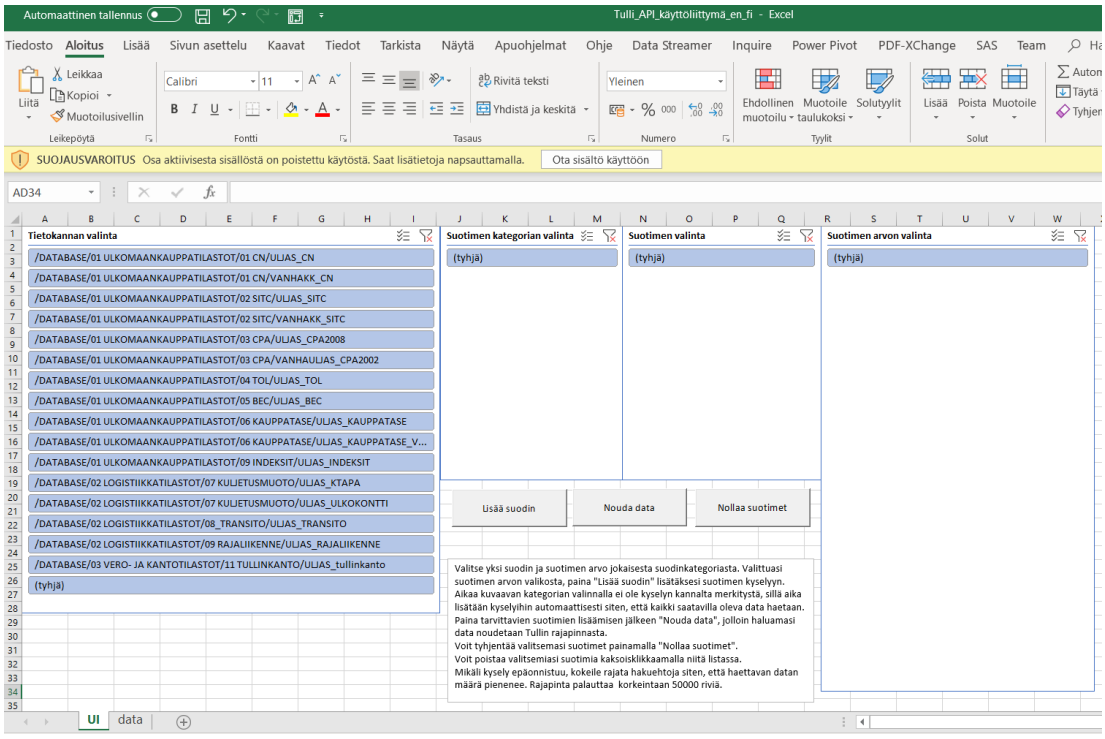
Excel-tiedostoon on rakennettu käyttöliittymä, jonka avulla haetaan reaaliaikaisesti dataa Tullin rajapinnasta halutuilla parametreilla. Rajapintakyselyt tapahtuvat dynaamisesti käyttäjän valintojen perusteella. Käyttöliittymässä on nähtävissä valitun tietokannan tukemat kyselytyypit ja parametrit, joista käyttäjä voi valita haluamansa. Haetusta datasta piirretään automaattisesti kuvaaja. Käyttäjä voi jatkokäsitellä dataa helposti kopiomalla noudetun datan Excelissä joko manuaalisesti, tai käyttämällä data-sivun kopiointipainikkeita.

Käyttäjän täytyy avatessaan tiedostoa antaa Excelille lupa makrojen suorittamiseen ja ulkopuolisen sisällön käyttöön, koska tiedoston toiminnallisuus on näistä ominaisuuksista riippuvainen. Tämä tapahtuu klikkaamalla painiketta `Ota sisältö käyttöön` näytön yläreunaan avautuvassa suojausvaroituksesta kertovassa ilmoituksessa (Kuva 4).



Kuva 4. Suojausvaroitus

Excel-työkalu (Kuva 5) mahdollistaa tullidatan hakemisen eri tietokannoista (Liite 1). Tietokantakohtaisesti on käytössä erilaisia suotimia (eli parametreja), joiden valinta (Luku 3.1) vaikuttaa noudettavaan dataan, jonka jatkokäsittelyssä voidaan hyödyntää Excelin tai jonkin muun työkalun analytiikkaominaisuuksia (Luku 3.2). Luvussa 3.3 esitellään käytetyt teknologiat, luvussa 3.4 Excel-työkalun toiminnallisuus ja luvussa 3.5 jatkokehityskohteet.



Kuva 5. Excel-työkalu

Joissakin datakuutioissa on esiintynyt suorituskykyongelmia niiden sisältämän suuren suodatinvalikoiman vuoksi.

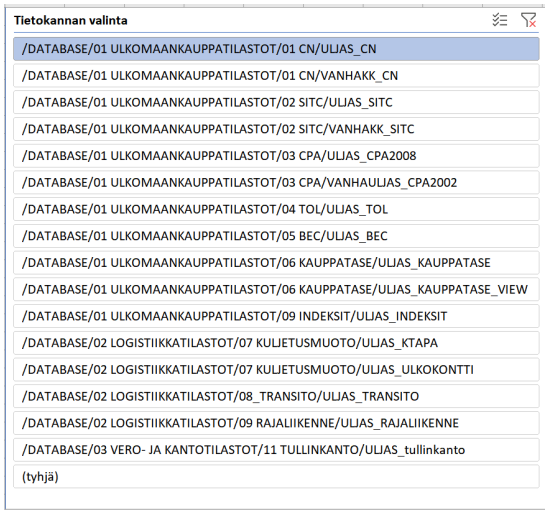
Suorituskykyongelmiin on pyritty löytämään erilaisia parannuksia ja ratkaisuja. Merkittävä parannus saatiin sillä, että suotimen arvolistan ensimmäisen arvon automaattinen valinta on otettu pois käytöstä. Joissakin datakuutioissa valittavia arvoja on noin 25000, jolloin ensimmäisen valinnan valitsemiseen käytetty menetelmä on liian hidas. Nopeampaa menetelmää ei ole selvitysten ja kokeilujen perusteella Excelissä saatavilla. Lisätty korvaava toiminnallisuus, joka tarkistaa onko käyttäjä valinnut listalta arvon tapahtuu kun `Lisää suodin` -painiketta painetaan.

3.1 Suotimien valinta

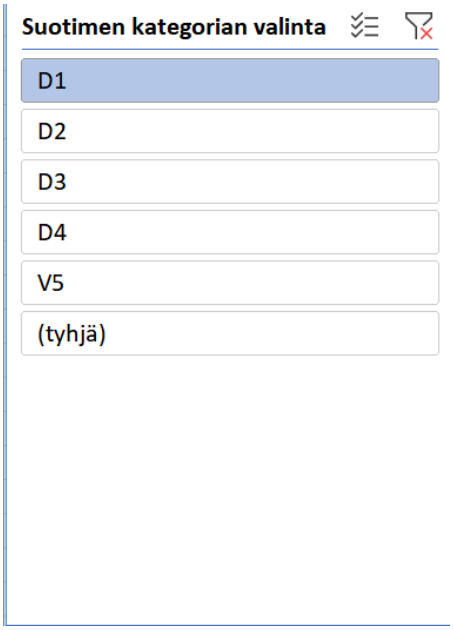
Noutaakseen dataa käyttäjän tulee valita haluamansa tietokanta `Tietokannan valinta` -valitsimella (Kuva 6), sekä yksi suodin ja suotimen arvo jokaisesta saatavilla olevasta kategoriasta. Valintojen tekeminen tapahtuu klikkaamalla hiirellä valinnat sisältäviä listoja. Listoissa esiintyy (tyhjä)-valintoja, jotka ovat käyttäjän valittavissa. Näitä valintoja ei tule käyttää – ne ovat listoissa mukana teknisistä syistä, eikä niitä Excelin ominaisuuksien vuoksi ole mahdollista poistaa. Mikäli käyttäjä valitsee (tyhjä)-valinnan, ohjelma varoittaa käyttäjää väärästä valinnasta ja peruuttaa valinnan.

Valittuaan `Tietokannan valinta` -listalta haluamansa tietokannan käyttäjä saa näkyviin tietokannan sisältämät suotimien kategoriat `Suotimen kategorian valinta` -listaan (Kuva 7), joista ensimmäinen valitaan automaattisesti. Ensimmäisen suodinkategorian sisältämät suotimet latautuvat `Suotimen valinta` -listaan (Kuva 8), joista vastaavasti ensimmäinen valitaan automaattisesti. Ensimmäisen suotimen sisältämät suotimen arvot latautuvat `Suotimen arvon valinta` -listaan (Kuva 9). Tässä tapauksessa ensimmäistä suotimen arvoa ei valita automaattisesti suorituskykysyistä. Suotimen arvon valinta ei myöskään tarjoa käyttäjälle lisää valintoja valitun arvon perusteella, kuten aikaisempien valintojen tapauksessa.

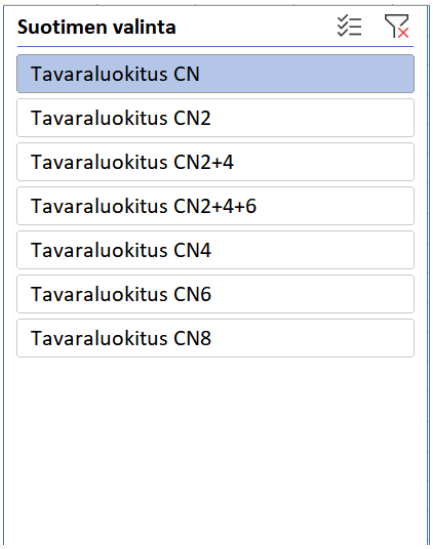
Käyttäjä voi valita vapaasti hiirellä haluamiensa tietokantojen, suodinkategorioiden, suotimien sekä suotimien arvojen välillä.



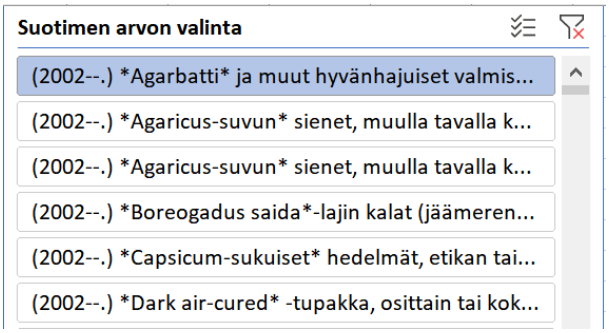
Kuva 6. Tietokannan valinta



Kuva 7. Suotimen kategorian valinta



Kuva 8. Suotimen valinta

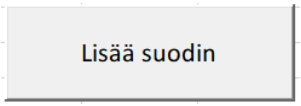


Kuva 9. Suotimen arvon valinta

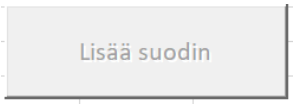
Valittuaan valitsemastaan tietokannasta haluamansa suodinkategorian, suotimen sekä suotimen arvon, käyttäjä voi lisätä suotimen aktiivisena oleviin suotimiin (Kuva 10) painamalla `Lisää suodin` -painiketta (Kuva 11). Suotimia valittaessa on huomionarvoista se, että aikaa kuvaava suodin toimii muista suotimista poikkeavasti. Teknisistä syistä aikaa kuvaava suodin täytyy valita, mutta valinnalla ei ole vaikutusta ohjelman toimintaan. Ohjelma noutaa aina saatavilla olevan datan mahdollisimman pitkältä aikaväliltä.

Valittuja suotimia voi poistaa listalta kaksoisklikkaamalla, ja korvaavan suotimen voi valita normaalisti käyttöliittymän kautta tässä luvussa kuvatun mukaisesti.

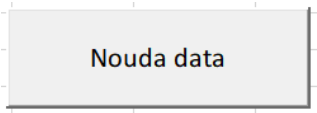
Kuva 10. Aktiiviset suotimet



Kuva 11. Lisää suodin -painike (aktiivinen)



Kuva 12. Lisää suodin -painike (inaktiivinen)



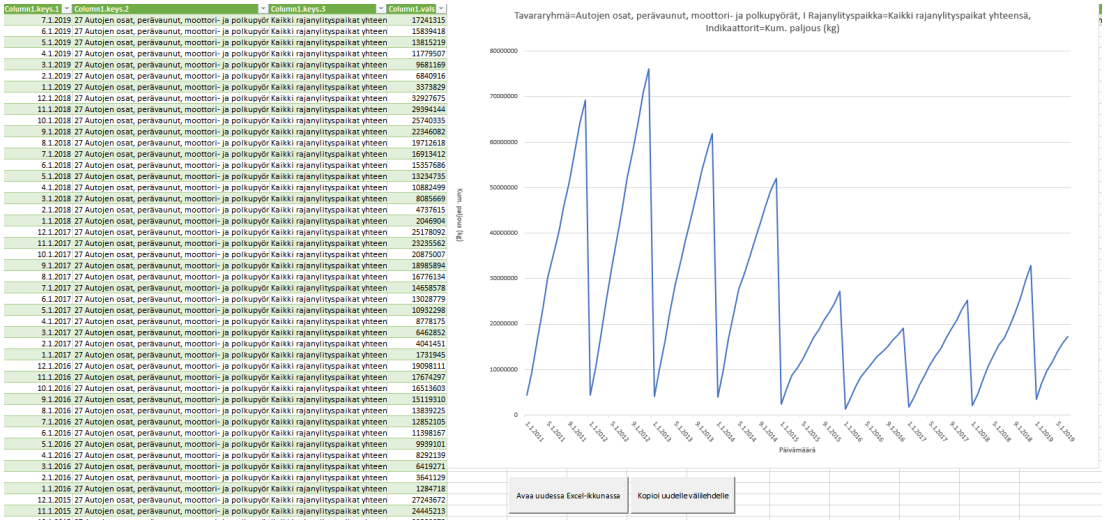
Kuva 13. Nouda data -painike

Käyttäjän valittua yhden suotimen ja suotimen arvon jokaisesta kategoriasta `Lisää suodin` -painikkeen teksti muuttuu harmaaksi ja painike poistuu käytöstä, ilmaisten käyttäjälle, että kaikki tarvittavat suotimet on valittu. Samanaikaisesti `Nouda data` -painikkeen (Kuva 13) teksti muuttuu harmaasta mustaksi ja se otetaan käyttöön. Käyttäjä voi tällöin noutaa Tullin Uljas-rajapinnasta valitsemastaan tietokannasta valitsemiensa suotimien mukaisen datan `Nouda data` -painiketta painamalla.

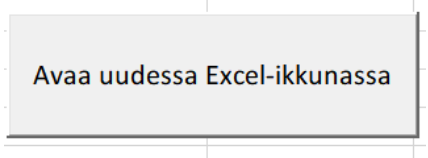
3.2 Data ja sen jatkokäsittely

Hakujen tulokset tallentuvat datataulukkoon omalle välilehdelleen `data` (Kuva 17), jossa on haetun datan lisäksi datan tulkinnan helpottamiseksi automaattisesti luotu viivakaavio. Näkymä siirtyy automaattisesti `data`-välilehdelle käyttäjän suoritettua haun Tullin Uljas-rajapinnasta. Käyttäjä voi siirtyä takaisin suodinvalintaan valitsemalla `UI`-välilehden (Kuva 17).

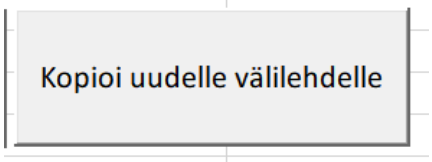
Mikäli käyttäjä haluaa tallentaa tai muokata datataulukkoa, on suositeltavaa käyttää välilehdeltä löytyvää `Avaa uudessa Excel-ikkunassa` -painiketta (Kuva 15). Tämä toiminto kopioi `data`-välilehden uuteen Excel-ikkunaan, jonka käyttäjä voi tallentaa muuttamatta alkuperäistä Excel-tiedostoa. Haetun datan välilehden voi myös halutessaan kopioida uudelle välilehdelle painamalla `Kopioi uudelle välilehdelle` -painiketta (Kuva 16), jolloin `data`-välilehdestä luodaan numeroitu kopio (Kuva 18: Kopioitu data-välilehti). Tässä tapauksessa datan käsittelystä aiheutuvat muutokset tallentuvat alkuperäiseen tiedostoon.



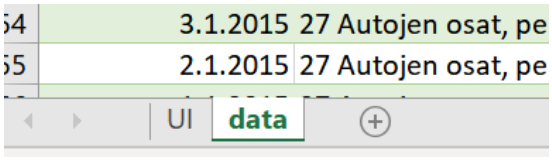
Kuva 14. Data-välilehti



Kuva 15. Avaa uudessa Excel-ikkunassa -painike



Kuva 16. Kopioi uudelle välilehdelle -painike



Kuva 17. Käyttöliittymän välilehdet

Kuva 18. Kopioitu data-välilehti

3.3. Käytetyt teknologiat

Tässä luvussa on kuvattu yleisellä tasolla työkalun toteutuksessa hyödynnetyt Microsoft Excel -taulukkolaskentaohjelman tukemat teknologiat. Luvussa 3.3.1 on kuvattu käytettyjä käyttöliittymäkomponentteja. Luvussa 3.3.2 kuvataan työkalun toteutuksessa käytetyt ohjelmointikielet.

3.3.1 Excel-taulukkolaskentaohjelman käyttöliittymäkomponentit

Työkalun toteutuksessa on käytetty useita erilaisia Excelin tarjoamia interaktiivisia käyttöliittymäkomponentteja, joihin on mahdollista kiinnittää toiminnallisuudeksi mielivaltaista VBA-ohjelmakoodia (Visual Basic for Applications, suom. Visual Basic sovelluksille). Esimerkiksi työkalun sisältämät painikkeet, kuten `Nouda data` -painike (Kuva 13), ovat Microsoft ActiveX-komponentteja, joita painaessa käynnistetään VBA-ohjelmakoodia sisältävä funktio.

Työkalun sisältämät valitsinlistat (esimeriksi `Tietokannan valinta`, Kuva 6) on toteutettu Pivot-taulukoita hyödyntäen. Pivot-taulukot on tarkoitettu datan yhteenvetojen laatimisen sekä jäsentämisen helpottamiseksi. Pivot-taulukkoon on mahdollista liittää Pivot-taulukon sisältöä käyttöliittymässä listamuodossa esittävä `Osittaja`, jonka avulla taulukon sisältöön voidaan vaikuttaa tekemällä valintoja listassa. Käyttöliittymän valintalistat on toteutettu tätä toiminnallisuutta hyödyntäen. Excelin tukemiin käyttöliittymäkomponentteihin kuuluu myös erilaisia kuvaajia – noudetusta datasta automaattisesti piirtyvä kuvaaja (Kuva 14) on toteutettu näitä komponentteja hyödyntäen. Kuvaaja on asetettu tarkkailemaan tiettyä kohtaa taulukosta, ja datan vaihtuessa kuvaajan käyttöliittymäkomponentti päivittyy kuvaamaan muuttunutta dataa.

3.3.2 Käytetyt ohjelmointikielet

Visual Basic for Applications (VBA) on useimpien Microsoftin Office -sovelluksien tukema ohjelmointikieli, jolla on mahdollista automatisoida moninaisia toimintoja Office-sovelluksien sisällä, sekä vaikuttaa Office-sovelluksien käyttöliittymän toimintaan. Työkalun julkaisuhetkellä viimeisin Visual Basic for Applications -ohjelmointikielen versio oli 7.1.

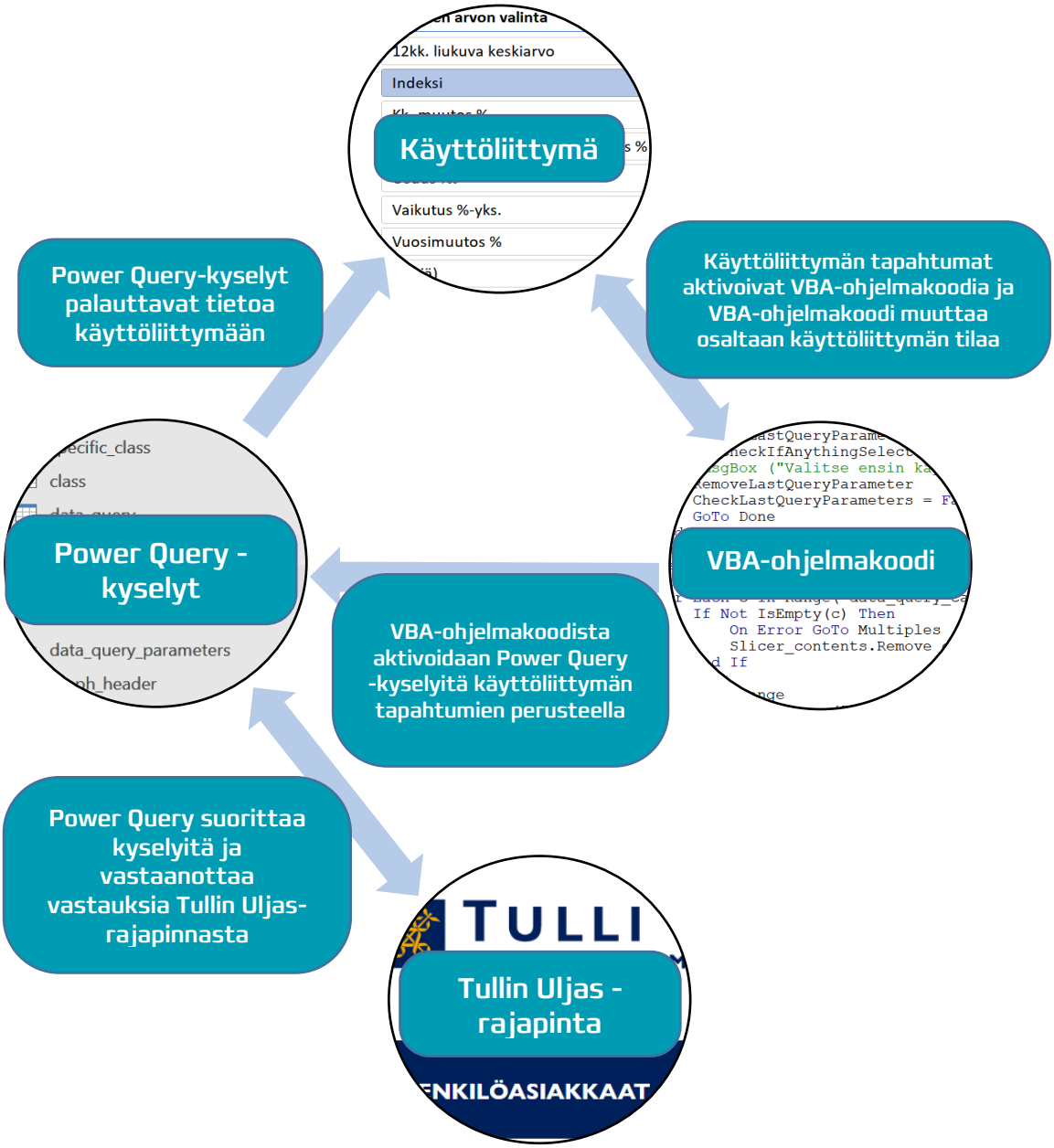
Power Query M formula language on Microsoft Power Queryn käyttämä ohjelmointikieli, joka on tarkoitettu datan käsittelyyn. Kielessä on toiminnallisuuksia datan noutamiseksi monista ulkopuolisista lähteistä sekä jäsentämiseksi käyttäjän tarpeiden mukaisesti.

3.4 Työkalun korkean tason toiminnallinen kuvaus

Työkalun korkean tason arkkitehtuuri perustuu Excelin sisältämien käyttöliittymäkomponenttien, taulujen, Visual Basic for Applicationsin, Power Queryn ja Tullin Uljas-rajapinnan yhteistoimintaan (Kuva 19).

Normaalissa käyttötapauksessa työkalu toimii yleisellä tasolla seuraavanlaisesti:

1. Haetaan Power Queryä hyödyntäen Tullin Uljas-rajapinnasta tieto rajapinnan sisältämistä tietokannoista (datakuutioista). Tieto tallennetaan Pivot-taulukkoon
2. Käyttäjä valitsee Pivot-taulukkoon kiinnitetystä `Osittajasta` haluamansa datakuution. Työkalu suorittaa dynaamisesti valinnan perusteella Power Query -kyselyn Tullin Uljas-rajapintaan, josta palautuu rajapinnasta tieto datakuution sisältämistä suodinkategorioista (dimensioista). Power Query-kysely huolehtii myös vastaanotetun datan muuntamisesta Exceliin tallentamiseksi sopivaan muotoon. Lopputuloksena syntyneet suodinkategoriat tallennetaan Pivot-tauluun. Suodinkategorioiden valitsimena toimivasta `Osittajasta` ensimmäinen dimensio valitaan automaattisesti Visual Basic for Applicationsia hyödyntäen, ja suoritetaan uusi Power Query kysely suodinkategorian sisältämistä suotimista. Vastaavasti suodinkategorioista valitaan ohjelmallisesti ensimmäinen, ja noudetaan Uljas-rajapinnasta saatavilla olevat suotimet. Tämän jälkeen ladataan suotimen mahdolliset arvot samalla tavoin kuin aikaisemmatkin tapaukset.
3. Käyttäjä valitsee haluamansa suotimet sekä suotimen arvot vaadituille suodinkategorioille (dimensioille) käyttöliittymää `(Osittajat` sekä `Painikkeet)` käyttäen. Valitut arvot tallennetaan työkalun taulukkoon niille varattuihin soluihin. Visual Basic for Applicationsia hyödyntäen valvotaan, ettei käyttäjä pääse tekemään vääriä valintoja, ja että tarvittavat asiat ovat tulleet valituksi.
4. Kun käyttäjä on valinnut ohjelmakoodissa tarpeelliseksi määritetyt suotimet, vapautetaan Visual Basic for Applicationsia käyttäen `Nouda data` -painike käytettäväksi. Vastaavasti `Lisää suodin` -painike lukitaan pois käytöstä.
5. Käyttäjä painaa `Nouda data` -painiketta. Painike suorittaa siihen liitetyn Visual Basic for Applications -funktion, joka puolestaan suorittaa valmiiksi laaditun Power Query -kyselyn datan noutamiseksi. Parametrit kyselyä varten ovat osittain kiinteästi määrättyjä (esim. datan muoto ja kieli), ja osittain ne poimitaan käyttäjän valintojen seurauksena taulukkoon tallennetuista arvoista. Power Query -kysely muuntaa taulukkoon tallennetut valinnat ja kiinteät parametrit rajapintakyselylle sopivaan muotoon.
6. Power Query -kysely palauttaa ja muotoilee vastaanottamansa datan Excelissä näytettäväksi soveltuvaan muotoon erilliselle välilehdelle (`data)`. Välilehdellä on kuvaaja, joka tarkkailee välilehden taulukkoa muutosten varalta. Kuvaaja päivittyy automaattisesti vastaamaan välilehden sisältämää dataa. Kuvaajan otsikko muodostetaan Visual Basic for Applicationsia hyödyntäen ohjelmallisesti osin käyttäjän tekemien valintojen, ja osin Tullin Uljas-rajapinnan palauttaman datan perusteella.



Kuva 19. Työkalun korkean tason riippuvuudet

3.5 Työkalun ei-toteutetut kehityskohteet

Useamman suotimen vapaasti valittavan arvon lataaminen vaatii usean erillisen kyselyn suorittamisen peräkkäin, jota varten täytyy tehdä oma toiminnallisuus. Esimerkiksi yhdellä kyselyllä saa karkeasti rajattua haun aikaväliä alusta tai lopusta, mutta sitä varten täytyy tehdä oma käyttöliittymä ja toiminnallisuus, mikäli sellaiselle on tarvetta.

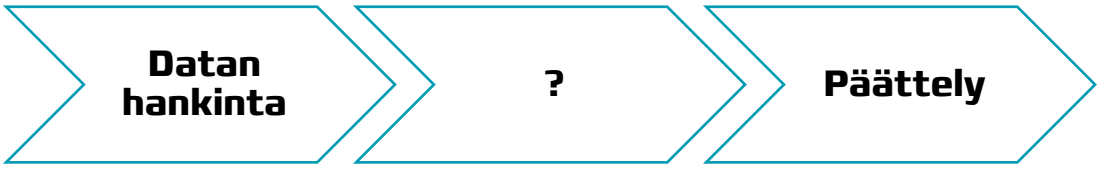
4 Colab-työkalu

Google Colaboratory -ympäristössä voidaan tehdä suoritettavia asiakirjoja. Ympäristö tarjoaa mahdollisuuden kokeilla erilaisia koneoppimisen (Kuva 20) ja automaattisen koneoppimisen (Kuva 21) paketteja kuten TPOT.

Tässä luvussa käsitellään Google Colaboratory -ympäristössä toteutetun Jupyter Notebook[7] suoritettavaan asiakirjastandardiin perustuvan työkalun toimintaa yleisellä tasolla. Työkalun korkean tason arkkitehtuurikuvaus on nähtävissä kuvassa 28.

Työkalun toiminnallisuudeksi rajattiin tässä yhteydessä aikasarjaennusteiden tuottaminen ja tuotettujen ennusteiden tarkkuuden mittaaminen. Työkalun suunnittelussa ja toteutuksessa pyrittiin ottamamaan huomioon mahdollisuuksien rajoissa helppokäyttöisyys ja uudelleenkäytettävyys – tavoitteena oli, että työkalusta olisi hyötyä myös tulevaisuudessa, ja että käyttötapaus ei rajoittuisi ainoastaan yhden datasetin analysointiin.

Kuva 20. Koneoppimisen työnkulku



Kuva 21. Automaattisen tai lisätyn koneoppimisen työnkulku

Piirresuunnittelussa hyödynnetään erilaisia muunnostekniikoita, jotka voidaan jakaa neljään ryhmään: normalisointi, enkoodaaminen, imputointi ja poikkeamien (outliers) käsittely. Taulukko 1 kuvaa käytettyjen työkalujen ominaisuuksia tämän toiminnallisuuskategorian osalta. Mikäli toiminnallisuus on ollut dokumentaatiosta ymmärrettävissä selvästi olemassa olevaksi, on taulukkoon merkitty työkalun kohdalle ”x”.

Taulukko 1. Normalisointi (kuten logaritmisointi ja skaalaus), enkoodaaminen (kuten one-hot), imputointi (kuten interpolointi).

Paketti	Normalisointi	Enkoodaaminen	Imputointi	Poikkeamat
TPOT	x		x	
Auto-sklearn	x	x	x	
Prophet	x		x	x
Pyramid-Arima				

[7] https://jupyter-notebook.readthedocs.io/en/stable/

Luvussa 4.1 käydään läpi yleisellä tasolla työkalun käyttöliittymän tavoitteita ja toimintaa. Luvussa 4.2 käsitellään aikasarjaennustamisessa käytettäviä automaattisen koneoppimisen paketteja/työkaluja. Lisäksi Colab-työkalulla on mahdollista tehdä regressioita ja luokittelijoita (Luku 4.3). Luvussa 4.4 käydään läpi työkalun tuottamat lopputuotteet (tiedostot). Luvussa 4.5 käydään lyhyesti läpi sitä, kuinka työkalun toimivuutta on testattu.

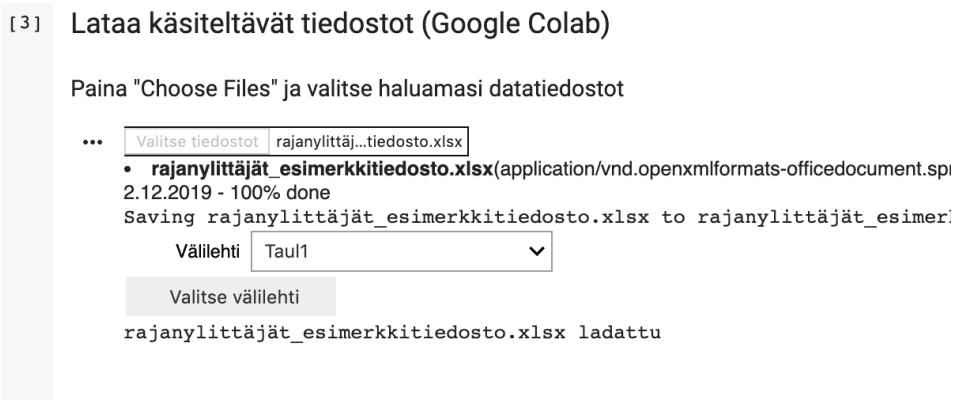
4.1 Käyttöliittymä

Työkalu on toteutettu Jupyter Notebook -standardin mukaisen suoritettavan asiakirjan avulla. Jupyter Notebook -asiakirjoissa on tyypillisenä ominaisuutena ohjelmakoodin sulkeminen toisistaan visuaalisesti erotettuihin soluihin, jotka suoritetaan toisistaan jossain määrin erillisinä kokonaisuuksina. Ohjelman nimiavaruus säilyy eri solujen suorituksien välillä, joten solujen välillä on mahdollista olla riippuvuuksia, joita tässäkin työkalussa on hyödynnetty.

Koska työkalun tavoitteisiin kuului helppokäyttöisyys ja uudelleenkäytettävyys, toteutettiin työkaluun hiirellä käytettävä käyttöliittymä, jonka avulla käyttäjän on mahdollista käyttää työkalua ilman, että käyttäjän tarvitsee ymmärtää tai nähdä toteutuksen matalamman tason yksityiskohtia, kuten esimerkiksi ohjelmakoodia.

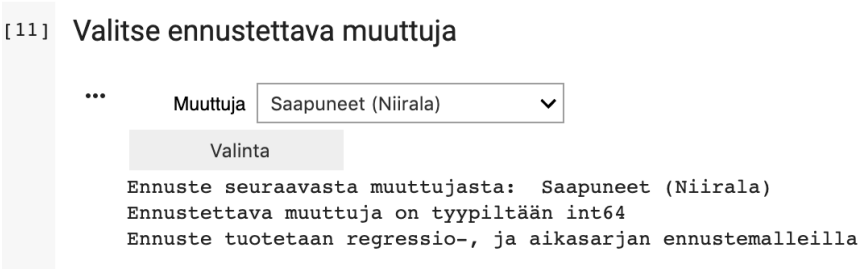
Alla on kuvattuna joitakin käyttöliittymään toteutettuja toimintoja – kaikkia toimintoja ei nähty tarkoituksenmukaiseksi kuvata tässä yhteydessä, sillä niiden kuvaukset löytyvät suoritettavasta asiakirjasta.

Käyttäjän on mahdollista ladata omalta koneeltaan haluansa aikasarjadataa sisältävä tiedosto analysoitavaksi suoritettavaan asiakirjaan suoraan käyttöliittymän kautta (Kuva 22).



Kuva 22. Tiedoston lataus työkaluun

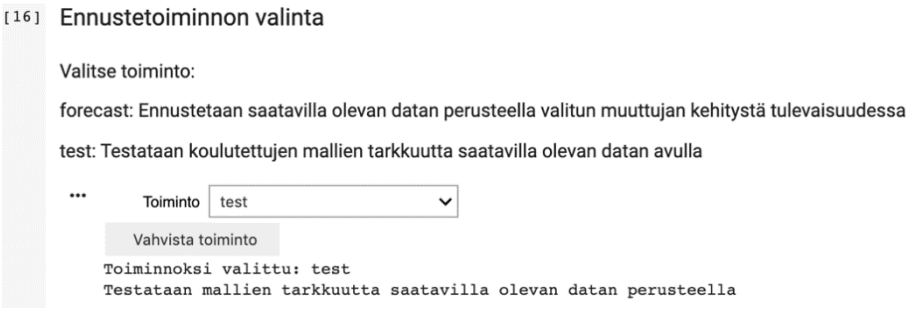
Käyttäjän on mahdollista valita lataamastaan tiedostosta ennustettava muuttuja (sarake) käyttöliittymän avulla. Työkalun mahdollisesti tulevaisuudessa ilmestyvän version on tarkoitus tukea myös regressio- ja luokittelijamalleja datalle, joten ennustettavan muuttujan tietotyypin perusteella valitaan käytettävä ennustemalli.



Kuva 23. Ennustettavan muuttujan valinta

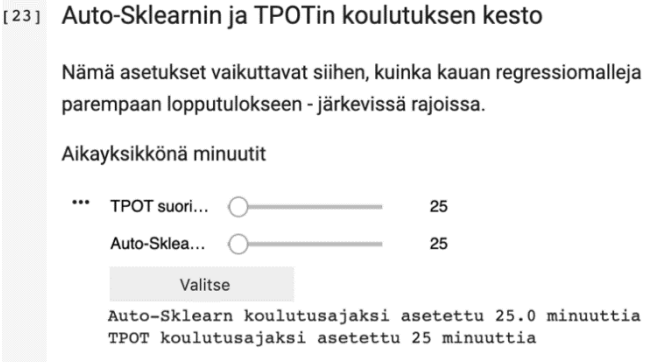
Käyttäjän on mahdollista valita käyttöliittymän välityksellä työkalun toimintatila. Työkalussa on kaksi erilaista toiminnallisuutta: ennusteen tuottaminen, sekä tuotetun ennusteen tarkkuuden testaus (Kuva 24).

Ennusteen tuottamiseen tarkoitetussa tilassa käytetään käyttäjän syöttämää aikasarjadataa kokonaisuudessaan mallien kouluttamiseen, jonka jälkeen koulutetut mallit tuottavat halutun pituisen ennusteen datasta. Ennusteen tarkkuuden mittaamiseen tarkoitetussa tilassa käyttäjän syöttämästä datasta vain osaa käytetään mallien kouluttamiseen. Loppupäätä datasta käytetään aiemman datan perusteella luotujen ennusteiden vertailukohtana, jolloin on mahdollista saada kuva käytettyjen mallien tuottaman ennusteen tarkkuudesta.



Kuva 24. Ennustetoiminnon valinta

Käyttäjän on myös mahdollista valita käyttöliittymän kautta joitakin käytettyjen työkalujen hyväksymiä parametreja, jotka vaikuttavat työkalujen toimintaan. Kuvassa 25 kuvataan, kuinka esimerkiksi työkalussa hyödynnettyjen TPOT- ja auto-sklearn-työkalujen parametreihin voidaan vaikuttaa. Käyttäjän on mahdollista valita työkalujen operaatioihinsa käyttämä sallittu maksimiaika.



Kuva 25. Koulutuksen keston valinta

4.2 Aikasarjadatan ennustaminen

Aikasarjaennusteiden tarkoituksena on tuottaa saatavilla olevasta mittausdatasta ennusteita tulevaisuuteen sijoittuvista mittauspisteiden arvoista. Ennusteiden tarkkuus riippuu sekä käytetyn datan ominaisuuksista, että ennusteen luomiseen käytetyistä malleista. Aikasarjanennusteiden toteuttamisen osalta valittiin menetelmiksi Prophet[8] (Luku 4.1.2), Pyramid-Arima[9] (Luku 4.2.1), sekä räätälöitynä toteutuksena ARIMA-pohjainen monimuuttujaregressiomalli (Luku 4.2.3).

4.2.1 Pyramid-Arima

Pyramid-Arima on paketti, jonka tavoitteena on tuoda R-ohjelmointikielen auto.arima-paketin toiminnallisuus Python-ohjelmointikielen käyttäjien saataville. Paketti hyödyntää Pythonin *statsmodels.tsa.ARIMA-* sekä *statsmodels.tsa.statespace.SARIMAX* -luokkia, yhdistäen näiden luokkien toiminnallisuuden yhden helppokäyttöisen työkalun alle. Paketin tukemalla SARIMAX-menetelmällä on mahdollista tuottaa myös ulkopuolisia muuttujia (exogenous regressors) huomioon ottavia ennusteita. [10]

4.2.2 Facebook Prophet

Facebook Prophet on Facebookin Core Data Science Teamin kehittämä aikasarjaennusteiden tuottamiseen tarkoitettu kirjasto. Se perustuu additiiviselle mallille, ja sen ominaisuuksissa on pyritty ottamaan huomioon erityisesti kausivaihtelun arviointi sekä esimerkiksi juhlapyhien ajankohdat. [11]

4.2.3 Monimuuttujaregressio ARIMA-ennusteista

Mikäli ennustettavan muuttujan ja datasettiin sisältyvien muiden muuttujien välillä on olemassa tilastollinen yhteys, on mahdollista tehdä ennusteita ennustettavan muuttujan mahdollisesti tulevaisuudessa saamista arvoista datasettiin sisältyviä muita muuttuja hyödyntäen.

Tätä varten työkaluun toteutettiin toiminallisuus, joka mahdollistaa aikasarjaennusteiden tuottamisen Pyramid arima -kirjastosta löytyvällä ARIMA-menetelmällä (auto_arima) aikasarjadataa sisältävän datasetin kaikista muuttujista. Koska aikasarjaennusteet ovat kustakin muuttujasta saatavilla, voidaan näiden aikasarjaennusteiden avulla arvioida monimuuttujaregression keinoin ennustettavan muuttujan arvoja.

Menetelmässä mahdollisina ennusteiden virheellisyyttä aiheuttavina tekijöinä ovat datasetin muuttujien mahdollinen vähäinen tilastollinen yhteys ennustettavaan muuttujaan, sekä muuttujista toteutettavien aikasarjaennusteiden epätarkkuus. Menetelmän käyttö edellyttää aikasarjaennusteiden tuottamista regressiomallien tuottamiseen käytetyistä muuttujista.

[8] https://facebook.github.io/prophet/
[9] https://pypi.org/project/pyramid-arima/
[10] Ibid, Pyramid Arima
[11] Ibid, Facebook Prophet

4.3 Regressio ja luokittelijat

Colab-toteutukseen valittiin automaattisen koneoppimisen työkaluksi TPOT[12], joka automatisoi muun muassa piirteiden valinnan ja hyperparametrien optimoinnin (Luku 4.3.1). Lisäksi valittiin toiseksi automaattisen koneoppimisen työkaluksi auto-sklearn. Se on yleisesti käytetyn sklearn-työkalun[13] (toolkit) laajennus ja se tuottaa muun muassa yhdistemalleja (ensemble models) (Luku 4.3.2).

4.3.1 TPOT

TPOT on avoimeen lähdekoodiin perustuva automaattiseen koneoppimiseen tarkoitettu ohjelmakirjasto. Se tukee useita eri koneoppimismalleja sekä muita datan käsittelyyn tarkoitettuja työkaluja. TPOT on rakennettu scikit-learn kirjaston toiminnallisuutta hyödyntäen. [14]

TPOTin tärkeimpiin ominaisuuksiin kuuluu koulutukseen käytettävän datan perusteella parhaan koneoppimismallin, sekä mallin hyperparametrien valinta automaattisesti. Tässä toiminnallisuudessa on hyödynnetty geneettisiä algoritmeja. [15]

4.3.2 Auto-sklearn

Auto-sklearn on avoimeen lähdekoodiin perustuva automaattisen koneoppimiseen tarkoitettu ohjelmakirjasto. Sen tarkoituksena on olla suora korvaava toiminnallisuus scikit-learn-kirjaston estimator-toiminnallisuudelle, jonka pääasiallisena tehtävänä on luoda koneoppimismalleja syötetyn datan perusteella. [16]

Auto-sklean hyödyntää bayesilaista optimointia, metaoppimista sekä niin sanottuja ensemble-malleja, joissa hyödynnetään kahta tai useampaa koneoppimismallia samanaikaisesti.[17]

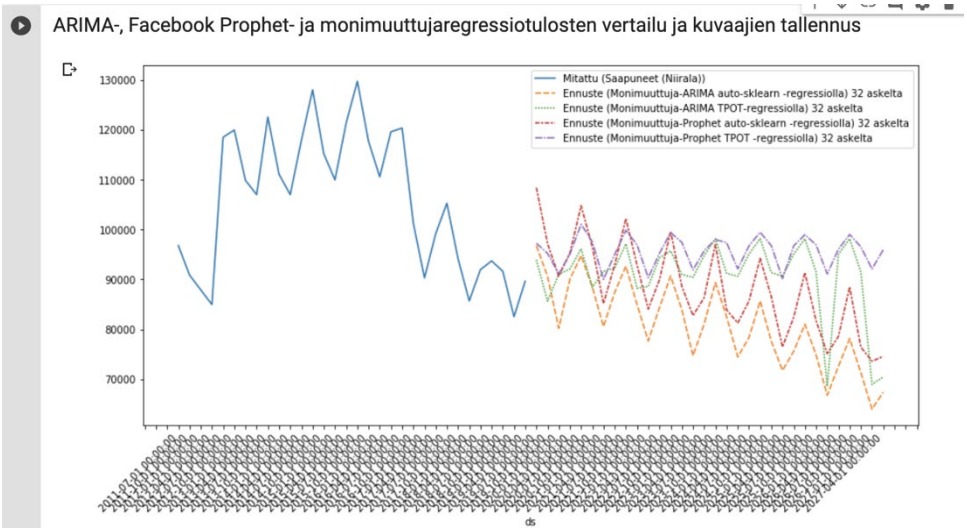
4.4 Työkalun tuottamat lopputuotteet

Työkalu tuottaa lopputuotteenaan useita erilaisia tiedostoja, joihin sisältyy työkalun eri toiminnallisuuksien lopputuloksia, sekä käyttäjän käyttöliittymän kautta valitsemia asetuksia, jotka vaikuttavat työkalussa hyödynnettyjen menetelmien tai työkalun itsensä toimintaan.

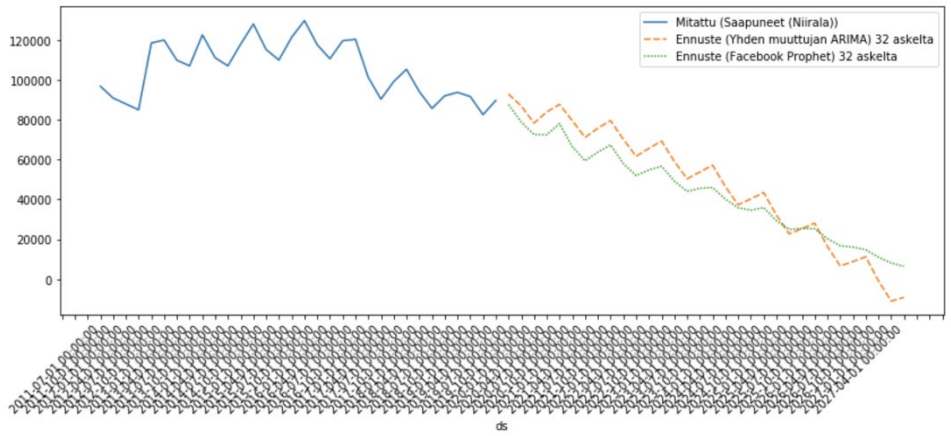
[12] https://epistasislab.github.io/tpot/
[13] https://automl.github.io/auto-sklearn/master/
[14] Ibid, TPOT
[15] Ibid, TPOT
[16] Ibid, auto-sklearn
[17] Ibid, auto-sklearn

Taulukko 2. Työkalun tuottamat tiedostot

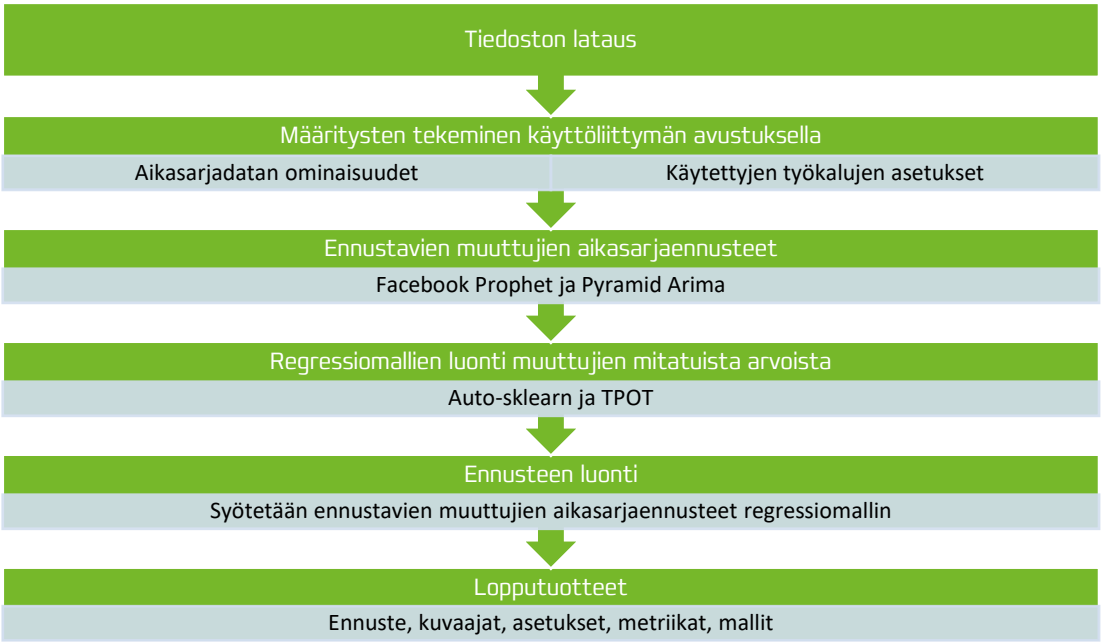
Tiedostonimi	**Kuvaus**
ennuste.xlsx	Työkalun tuottamat ennusteet excel-muodossa
mvarima_autosklearn.pickle	ARIMA-ennusteista tuotetun auto-sklearn regressiomallin tallennus
mvarima_tpot.pickle	ARIMA-ennusteista tuotetun TPOT-regressiomallin tallennus
plot1.png	Monimuuttujamallien tuloksista tuotettu kuvaaja (Kuva 26)
plot2.png	Yhden muuttujan mallien tuloksista tuotettu kuvaaja (Kuva 27)
prophet_tpot.pickle	Prophet-ennusteista tuotetun TPOT-regressiomallin tallennus
prophet_autosklearn.pickle	Prophet-ennusteista tuotetun auto-sklearn regressiomallin tallennus
stats.json	Tietoja työkalun tuottamien ennusteiden tarkkuudesta sekä käytetyistä asetuksista



Kuva 26. Esimerkki työkalun tuottamasta kuvaajasta monimuuttujaennusteiden osalta



Kuva 27. Esimerkki työkalun tuottamasta kuvaajasta yhden muuttujan ennusteiden osalta



Kuva 28. Colab-työkalun korkean tason arkkitehtuurikuvaus

4.5 Työkalun testaus

Tässä luvussa kuvataan työkalun toiminnallisuuden ja tarkkuuden testausta. Luvussa 4.5.1 on kuvattu, kuinka työkalun toiminnallisuutta on testattu, mitä virheitä testauksen aikana ilmeni ja mitä virheille tehtiin. Luvussa 4.5.2 on testattu työkalun tuottamien ennusteiden tarkkuutta vertailemalla niitä mitattuihin arvoihin.

4.5.1 Toiminnallisuuden testaus

Työkalu testattiin tämä dokumentin yhteydessä julkaistavalla *rajanylittäjät_esimerkkitiedosto.xlsx*-tiedostolla sekä Beijingin saasteisuutta ja siihen liittyviä muuttujia kuvaavalla 43284 riviä sisältävällä *pollution_beijing.xlsx*-aikasarjatiedostolla. Tiedoston sarakkeet on kuvattu Taulukossa 3.

Taulukko 3. Beijing-datasetin muuttujat (mukaillen https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/)

Sarakkeen otsikko	Kuvaus
No	Rivinumero
year	Vuosi
month	Kuukausi
day	Päivä
hour	Tunti
pm2.5	PM2.5-pitoisuus
DEWP	Kastepiste
TEMP	Lämpötila
PRES	Ilmanpaine
cbwd	Yhdistetty tuulen suunta
Iws	Kumulatiivinen tuulen nopeus
Is	Kumulatiivinen lumisateen kesto
Ir	Kumulatiivinen vesisateen kesto

Beijingin saasteisutta kuvaava *pollution_beijing.xlsx* datasetti käsiteltiin ennen työkaluun syöttämistä ensin Excelissä, jotta vuosia, kuukausia, päiviä ja tunteja kuvaavat erilliset sarakkeet saatiin yhdistettyä. Näillä toimenpiteillä ei ole työkalun teknisen toiminnan kannalta erityistä merkitystä – toimenpide tehtiin testaustyön helpottamiseksi, sillä näin aikasarjan aikaleimat saatiin sisältymään yhdelle sarakkeelle ja aikaleimojen käsittelyssä ilmenevät mahdolliset ongelmat ilmenisivät helpommin.

pollution_beijing.xlsx-tiedostoa testatessa ilmeni työkalun toiminnassa ongelma kategoristen muuttujien suhteen (tämän datasetin tapauksessa tuulen suunta), joka korjattiin muuntamalla kategoriset muuttujat numeeriseen muotoon.

Myöskin työkalun sijainnista Colab-ympäristössä johtuen ilmeni ongelmia Colabin ominaisuuksien rajoitteiden suhteen. Colabin käyttöliittymässä on aikaraja, joka katkaisee yhteyden suoritettavaan asiakirjaan, mikäli käyttöliittymässä ei tehdä mitään Colabin aktiivisuudeksi luokiteltavaa toimintaa – eli esimerkiksi mikäli käyttäjä jättää koneoppimisalgoritmin Colabiin pyörimään, ja poistuu itse paikalta. Colabin suoritus jatkuu käyttöliittymäyhteyden katkeamisesta huolimatta taustalla, mutta tulosteiden saaminen käyttöliittymästä muuttuu mahdottomaksi. Aikarajoite teki koneoppimismallien pidemmästä kouluttamisesta haastavaa, mikä on epätoivottavaa, sillä koneoppimismallit yleisesti ottaen hyötyvät Colabin sallimaa aikarajaa pidemmästä koulutusajasta.

Tämän ongelman ratkaisuksi toteutettiin työkalun tuottamien ennusteiden valmistuttua tulostettavia tiedostoja, joihin tallentuvat kaikki ajon kannalta olennaiset lopputuotteet, kuten tulokset, metriikat, kuvaajat ja asetukset.

4.5.2 Tarkkuuden testaus

Työkalun tarkkuutta testattiin *rajanylittäjät_esimerkkitiedosto.xlsx*-tiedoston avulla. Tiedostosta valittiin ennustettavaksi muuttujaksi Niiralan rajalta Suomeen tulijat. Tiedosto sisältää Niiralan rajanylittäjien lisäksi suuren määrän erilaisia muuttujia, joilla on katsottu olevan jollakin tavalla yhteydessä ennustettavaan muuttujaan. Muuttujiin kuuluu esimerkiksi kauppatase, Venäjältä Suomen kanssa maantieteellisesti läheisiin maihin rajanylittäjät, sekä muita vastaavia muuttujia. Tarkka listaus kaikista käytetyistä muuttujista löytyy Liitteestä 2.

Niiralan rajanylittäjien ennustamisessa (forecasting) työkalussa käytetyt autoML-työkalut antavat erilaisia tuloksia, joiden hyvyyteen liittyviä metriikoita voidaan verrata siinä tapauksessa, että ennustaminen aloitetaan menneisyydestä (backtesting). Taulukko 4 on kuvattuna Niiralasta saapuvista rajanylittäjistä eri menetelmin laadittujen ennusteiden hyvyys useilla yleisesti käytetyillä metriikoilla. Taulukko 4 kuvatut metriikat on luotu työkalun ajosta, jossa sekä TPOT- että auto-sklearn-oppijoiden koulutusajaksi oli asetettu 60 minuuttia. Molempia algoritmeja koulutetaan ajon aikana kahdesti, jolloin kokonaiskoulutusajaksi muodostuu 240 minuuttia. Lisäksi ajon, josta metriikat on muodostettu, alkuperäisestä datasta muodostettavien koulutus- ja testausjoukkojen jakosuhde oli asetettu arvoon 0.9 – eli 90 % datasta käytettiin koulutukseen, ja 10 % ennusteiden hyvyyden mittaamiseen.

Tarkemmat tiedot työkalun asetuksista löytyvät ajon lopussa syntyvästä stats.json-tiedostosta, jonka sisältö on kuvattuna Liitteessä 2.

Metriikoiden laskemiseen on käytetty seuraavia sklearnin tarjoamia toiminnallisuuksia:
R2 score: *sklearn.metrics.r2_score()*
Mean absolute error: *sklearn.metrics.mean_absolute_error()*
Median absolute error: *sklearn.metrics.median_absolute_error()*
Mean squared error: *sklearn.metrics.mean_squared_error()*

Taulukko 4. Ennusteiden hyvyyden metriikat Niiralan rajanylittäjien ennusteelle

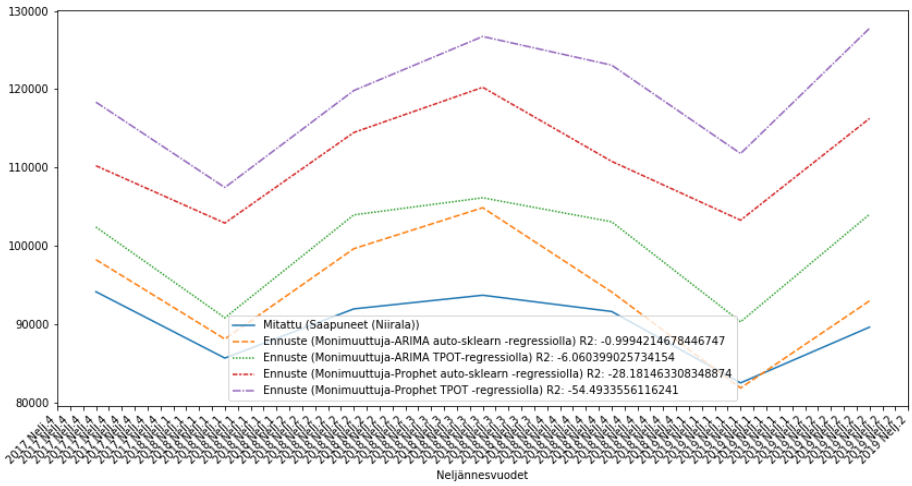
	Mean Absolute Error	R^2 score	Median Absolute Error	Mean Squared Error
Prophet	33546,1	-71,2787	33306,4	1158735024,7
Pyramid-Arima	7860,6	-4,2655	5369,5	84413489,8
Arima ulkopuolisilla regressoreilla (Arima)	16359,4	-25,5561	14841,5	425733458,5
Arima ulkopuolisilla regressoreilla (Prophet)	86758	-501,8512	88533	8061448751,3
Monimuuttujaregressio (TPOT)	10210,6	-6,0604	11448,2	113188641,2
Monimuuttujaregressio (auto-sklearn)	4554,8	-0,9994	3353,9	32053684

Kuva 29 on kuvattuna työkalun tuottama kuvaaja monimuuttujamenetelmällä tuotetuista ennusteista sekä vertailukohtana olevasta mitattujen arvojen sarjasta. Työkalu on kuvaajasta päätellen ennustanut säännönmukaisesti tulijoiden määrän mitattua oikeaa arvoa suuremmaksi. Tämän voisi olettaa viittavan siihen, että regressiomalliin mukaan otetut muuttujat – muuttujien suuresta määrästä huolimatta – eivät riitä kuvaamaan rajanylittäjien arvosarjan loppupäässä tapahtuvaa suhteellisen voimakasta laskua. Myös automaattiseen koneoppimiseen perustuvan mallin luomiseen välttämättä liittyvä satunnaisuus on otettava tässä yhteydessä huomioon – käytetyt automaattisen koneoppimisen työkalut eivät ole välttämättä löytäneet esikäsittely- ja ennustemalliensa mahdollisista eri permutaatioista juuri sitä yhdistelmää, jolla saataisiin kuvattua parhaiten ennusteeseen valittujen ennustavien muuttujien yhteys ennustettavaan muuttujaan.

Todennäköisimmältä vaikuttava syy epätarkkuudelle on kuitenkin se, että ennuste on toteutettu yksittäisten muuttujien aikasarjaennusteiden regressioista. Mikäli aikasarjaennusteet ovat jollakin tavalla epätarkkoja – jossakin ennustevoimaltaan mallin hyväksi määrittelemässä muuttujassa on tapahtunut esimerkiksi voimakas aikasarjaennustemallin osalta ennakoimatta jätetty trendimuutos ennusteaikana – lisääntyy näiden aikasarjaennusteiden pohjalta muodostetussa regressiomalliin perustuvassa ennusteessa epätarkkuus välttämättä.

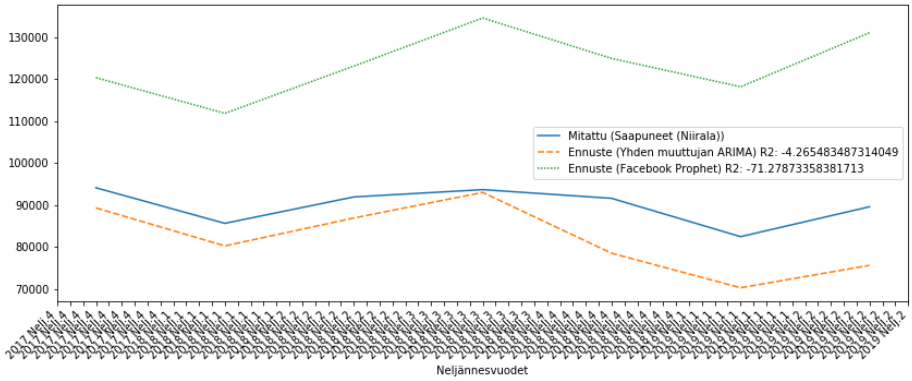
Kuitenkin kaikissa ennusteissa kausivaihtelu on taltioitunut malliin suhteellisen onnistuneesti, vaikkakaan kaikki ennusteet eivät ole suuruusluokaltaan onnistuneet Niiralan rajanylittäjien ennustamisessa.

Kuvassa 29 on huomionarvoista monen muuttujan Arima-ennusteeseen ja auto-sklearn regressiomalliin perustuvan ennusteen onnistuneisuus muihin työkalupermutaatioihin nähden – tämä ilmiö oli havaittavissa useimmilla suorituskerroilla tällä datalla. Tämä alleviivaakin useiden eri ennustemallivaihtoehtojen ja testauksen tärkeyttä – ei ole takeita siitä, että sama ennustetyökalujen yhdistelmä toimii jokaisella datasetillä yhtä hyvin, vaan muilla dataseteillä saattaa parhaan mallin luova yhdistelmä olla erilainen.



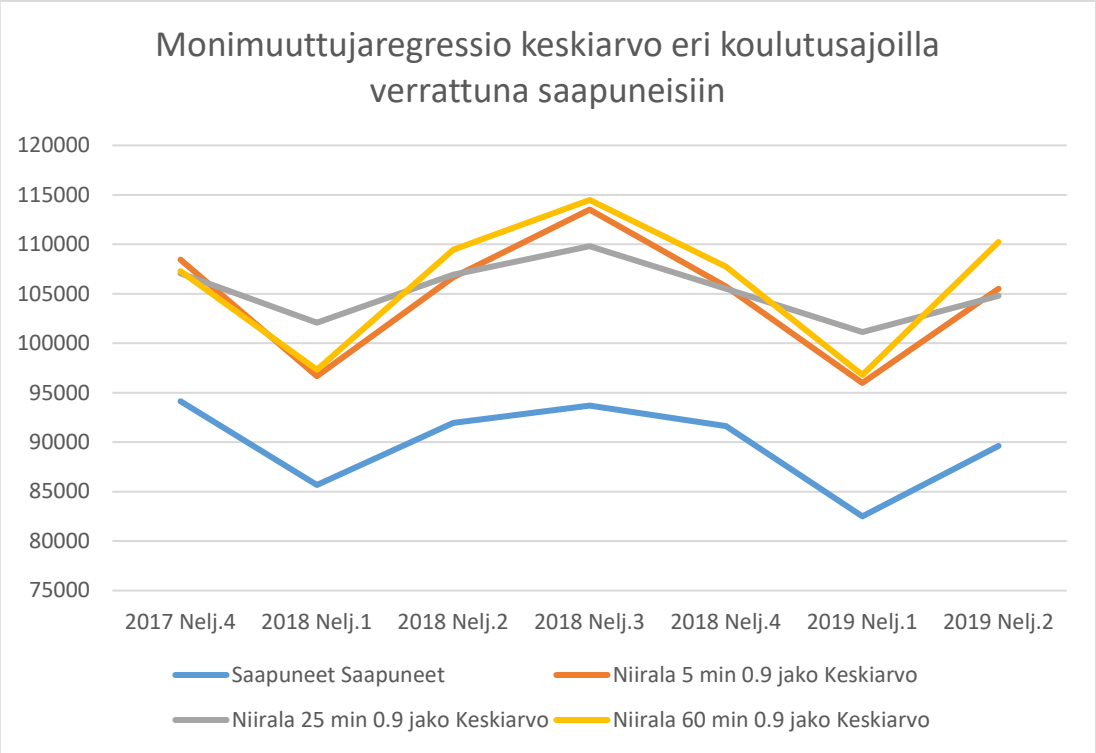
Kuva 29. Monimuuttujamenetelmien tuottamien ennusteiden vertaaminen mitattuihin arvoihin

Kuvassa 30 on kuvattuna yhden muuttujan aikasarjan ennustemallien tulokset tältä työkalun suorituskerralta. Kuvan perusteella Pyramid-Arima näyttää suoriutuneen ennusteesta Facebook Prophetia paremmin. Alkuperäistä dataa tarkastellessa Facebook Prophet on jättänyt huomiotta datan loppupäässä alkaneen laskutrendin, ja on muodostanut kausivaihtelulla varustetulta trendiviivalta näyttävän ennusteen. Tästä johtuen ennuste ei ole edes ennusteen alkupäässä yhtenevä mitatun datan kanssa. Pyramid-Arima on ottanut huomioon datan loppupäässä ilmenneen trendimuutoksen rajanylittäjien määrässä, ja on onnistunut aloittamaan ennusteensa suhteellisen läheltä mitattua määrää.



Kuva 30. Yhden muuttujan menetelmien tuottamien ennusteiden vertaaminen mitattuihin arvoihin

Kuvassa 31 monimuuttujaregression keskiarvo eri koulutusajoilla verrattuna saapuneisiin on kuvattuna muuten samoilla asetuksilla, mutta eri regressiomallien koulutusajoilla tuotettujen monimuuttujaennusteiden keskiarvot. Kuvassa on huomionarvoista, että 5 minuutin ja 60 minuutin koulutusajoilla toteutuneet ennusteiden keskiarvot ovat lähempänä toisiaan kuin 25 minuutin koulutusajalla toteutuneiden ennusteiden keskiarvot, toisin kuin saattaisi olettaa. Kuvasta on havaittavissa hyvin automaattiseen koneoppimiseen liittyvä satunnaisuus - 5 minuutin koulutuksella saatiin tässä tapauksessa keskimäärin lähes samanlainen ennuste kuin 60 minuutin koulutuksella, kun taas 25 minuutin koulutus, ollen koulutusajaltaan näiden kahden ääripään välissä, näyttää hyvinkin erilaiselta. 25 minuutin koulutusaika näyttäisi tuottaneen tässä tapauksessa tasaisimman ennusteen, joka osuu silmämääräisesti kausivaihtelultaan ennusteista lähimmäksi oikeaa arvoa. Kaikista ennusteista on kuitenkin havaittavissa jo aikaisempien kuvien kohdalla mainittu seikka, eli ennusteiden taipumus yliarvioida saapujien määrää.



Kuva 31. Monimuuttujaregression keskiarvo eri koulutusajoilla verrattuna saapuneisiin

Mallien tarkkuutta olisi todennäköisesti tulevaisuudessa mahdollista parantaa esimerkiksi ottamalla mallien toteutuksessa huomioon mahdolliset viiveet muuttujien riippuvuuksissa, mitä tässä julkaisussa kuvattu toteutus ei tee. Tämä ja vastaavat kehitysideat jäivät kuitenkin tämän version osalta toteuttamatta aikataulurajoitteiden vuoksi.

5 Niiralan rajanylittäjien ennustaminen

Kun halutaan ennustaa Niiralan rajaylittäjien määrää, Tulli-datasta otetaan vuosilta 2004-2019 Niiralaan saapuneet ja sieltä lähteneet, tuonti- ja vientimäärät, kauppatasetiedot, saapuneet ja lähteneet Norjaan ja Ruotsiin. Lisäksi Venäjän turvallisuuspalvelun avoimesta datasta[18] haettiin kuljetusmuodoittain (auto, ilmailu, juna, vesi, jalan) ja matkan syyn (liiketoiminta, työ, turismi, yksityinen, opiskelu, pysyvä muutto, ajoneuvon huoltohenkilökunta, sotilaat) mukaan eritellyt Venäjältä lähteneet Suomeen ja muihin Suomea maantieteellisesti lähellä oleviin maihin (Norja, Valko-Venäjä, Latvia, Liettua ja Viro). Ruplan kurssi haettiin OFX:n sivulta[19].

Luvussa 5.1 käydään läpi Niiralan rajanylittäjien määrän sekä joidenkin taloudellisten muuttujien kehitystä datasetin mittausvälillä kontekstina tuotetuille ennusteille ja löydetyille riippuvuustekijöille.

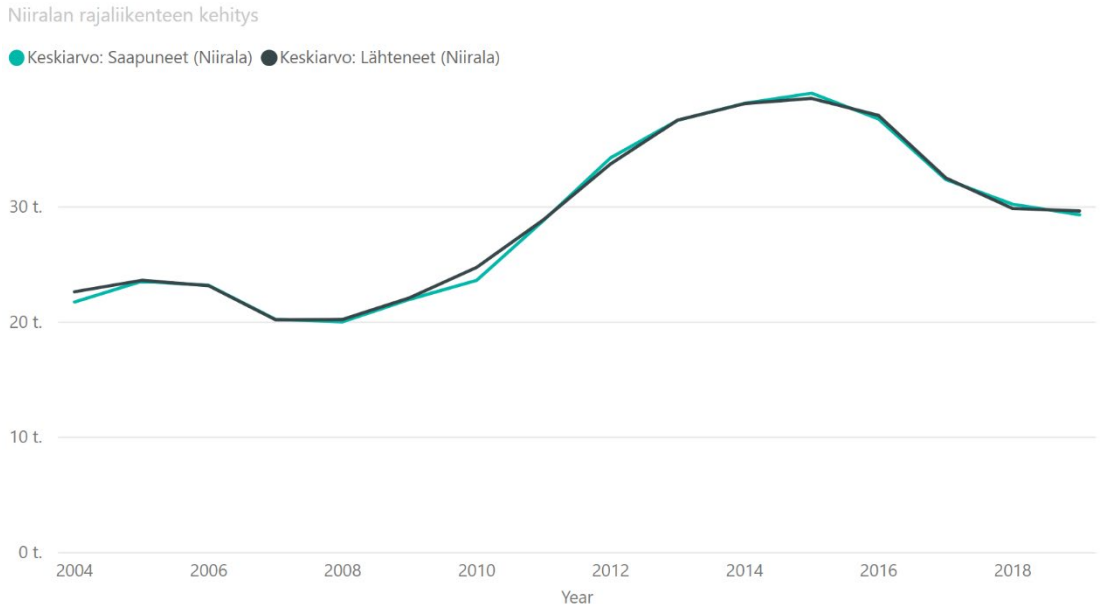
Luvussa 5.2 tarkastellaan Microsoft PowerBI:llä sekä sen tukemalla R-kielellä toteutetun analyysin tuloksia sen osalta, millaisia riippuvuuksia datasetissä esiintyvillä muuttujilla on toisiinsa – ymmärrettävästi etenkin datasetissä esiintyvien ennustavien muuttujien riippuvuussuhdetta Niiralan rajanylittäjien määrään.

Niiralan rajanylittäjien määrää ennakoivien ennusteiden tuottaminen perustuu suurelta osin projektin yhteydessä toteutetun Colab-työkalun hyödyntämiseen, jonka tuottamat tulokset ja tuloksista tehdyt päätelmät ovat luvussa 5.3.

5.1 Tarkasteluvälillä tapahtuneen kehityksen tarkastelu

Niiralan rajanylittäjien määrä saavutti huippuarvonsa noin vuonna 2015, kuten kuvasta 32 ilmenee. Rajanylittäjien määrä on ollut tarkastelujaksolla nousujohteinen, mutta vuoden 2015 jälkeen rajanylittäjien määrässä on tapahtunut selvää laskua – joskin määrä on tarkastelujakson lopussa (2019) vielä huomattavasti tarkastelujakson alhaisinta vuotta (2008) korkeampi. Saapuneiden ja lähteneiden määrät rajaliikenteessä seuraavat hyvin lähellä toisiaan koko tarkastelujaksolla.

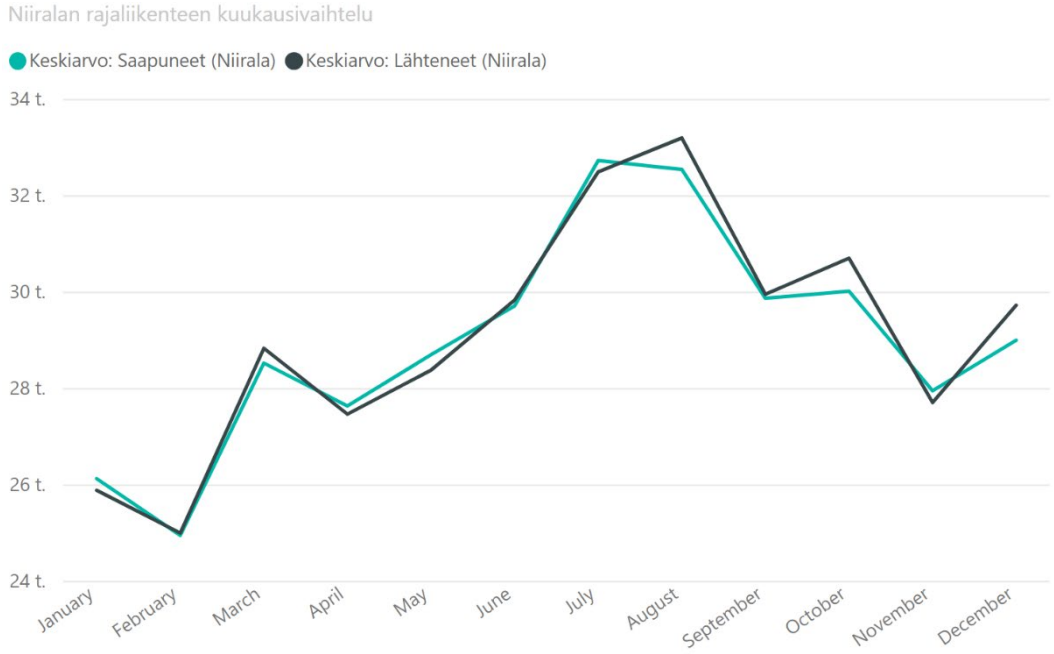
[18] https://fedstat.ru/indicator/38479
[19] https://www.ofx.com/en-au/forex-news/historical-exchange-rates/monthly-average-rates/



Kuva 32. Niiralan rajanylittäjien määrä 2004–2019

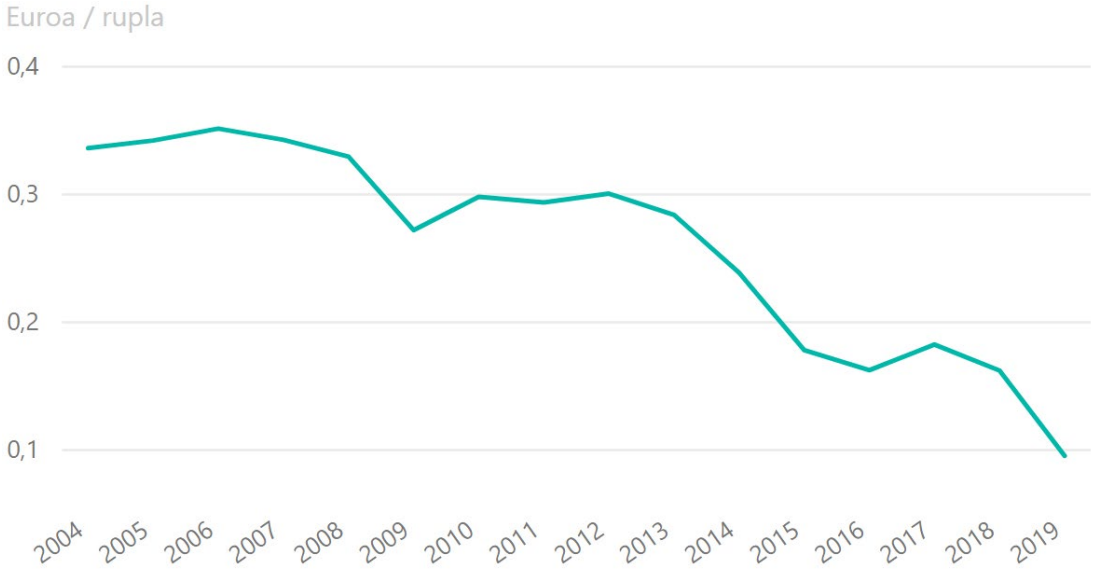
Kuten kuvasta 33 ilmenee, Niiralan rajanylittäjien määrän keskimääräinen huippukohta tarkastelujaksolla 2004 – 2019 asettuu selvästi elokuulle. Tämän voisi olettaa olevan seurausta kesälomista, jolloin ihmisillä on paremmin aikaa matkustaa. Pääsääntöisesti voidaan tehdä havainto siitä, että rajaliikenteen määrä kasvaa suhteellisen tasaisesti vuoden alusta elokuuhun, jonka jälkeen rajaliikenne alkaa vuoden loppua kohti keskimäärin laskea. Vuoden lopussa on nähtävissä pieni sesonki, joka on luultavasti joulun ja uudenvuoden seurausta.

Keskiarvoisessa tarkastelussa näyttäisi siltä, että rajaliikenteessä etenkin keskimääräisissä huippuarvoissa esiintyy lähtijöiden ja saapujien määrän välillä pientä eroa. Rajaliikenteen huippukohtina on lähteneiden määrä ollut pääsääntöisesti hieman saapuneiden määrää korkeampi.



Kuva 33. Niiralan rajanylittäjien kuukausittainen keskiarvo ajalta 2004–2019

Kuvassa 34 on kuvattuna Venäjän ruplan kurssi tarkastelujaksolla. Kurssi on ollut suhteellisen tasaisessa laskutrendissä lähes koko tarkastelujakson ajan, joskin kuvaajassa on havaittavissa joitakin paikallisia lasku- ja nousuhdanteita. Merkittävimmät paikalliset laskusuhdanteet ovat suhteessa nousuhdanteita jyrkempiä, ja näyttävät osuneen noin vuosille 2009 ja 2016. Merkittävimmät paikalliset nousuhdanteet näyttäytyvät vuosina 2010–2013 ja vuoden 2016 loppupuolelta vuoden 2017 loppupuolelle.

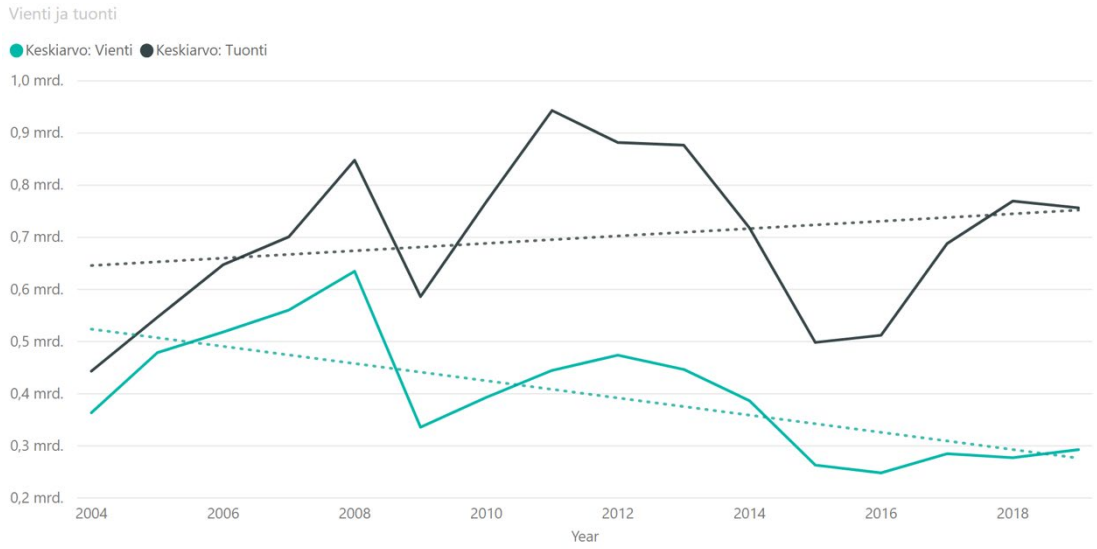


Kuva 34. Ruplan kurssi euroa / rupla

Kuten kuvasta 35 ilmenee, on tarkastelujaksolla tapahtunut Suomen ja Venäjän välisessä viennissä ja tuonnissa merkittäviä muutoksia. Kuvaajissa on havaittavissa selviä paikallisia nousu- ja laskusuhdanteita, ja kuvaajien paikalliset lasku- ja nousuhdanteet noudattelevat pääsääntöisesti toisiaan. Keskimääräinen vienti

Suomesta Venäjälle on tarkastelujaksolla laskenut jossain määrin, siinä missä tuonti on kasvanut – eli kauppatase on muuttunut. Kun tarkastelujakson alkupäässä kauppatase oli melko lähellä tasapainoa, tarkastelujakson loppupäässä kauppatase oli muuttunut Suomen osalta lähes 0,5 miljardia euroa alijäämäiseksi.

Osasyyn kauppataseen muutoksille voisi kuvitella olevan euron ja ruplan vaihtokurssin kehitys tarkastelujaksolla – ruplan edullisuus suhteessa euroon edistää Venäjältä lähtöisin olevaa vientiä. Kuvasta 36 kuitenkin ilmenee, että ruplan kurssilla ja tuonnilla ei ole tarkastellun datan perusteella tilastollisesti merkittävää yhteyttä.



Kuva 35. Tuonti ja vienti Venäjälle

5.2 Muuttujien tilastolliset yhteydet tarkasteluvälillä

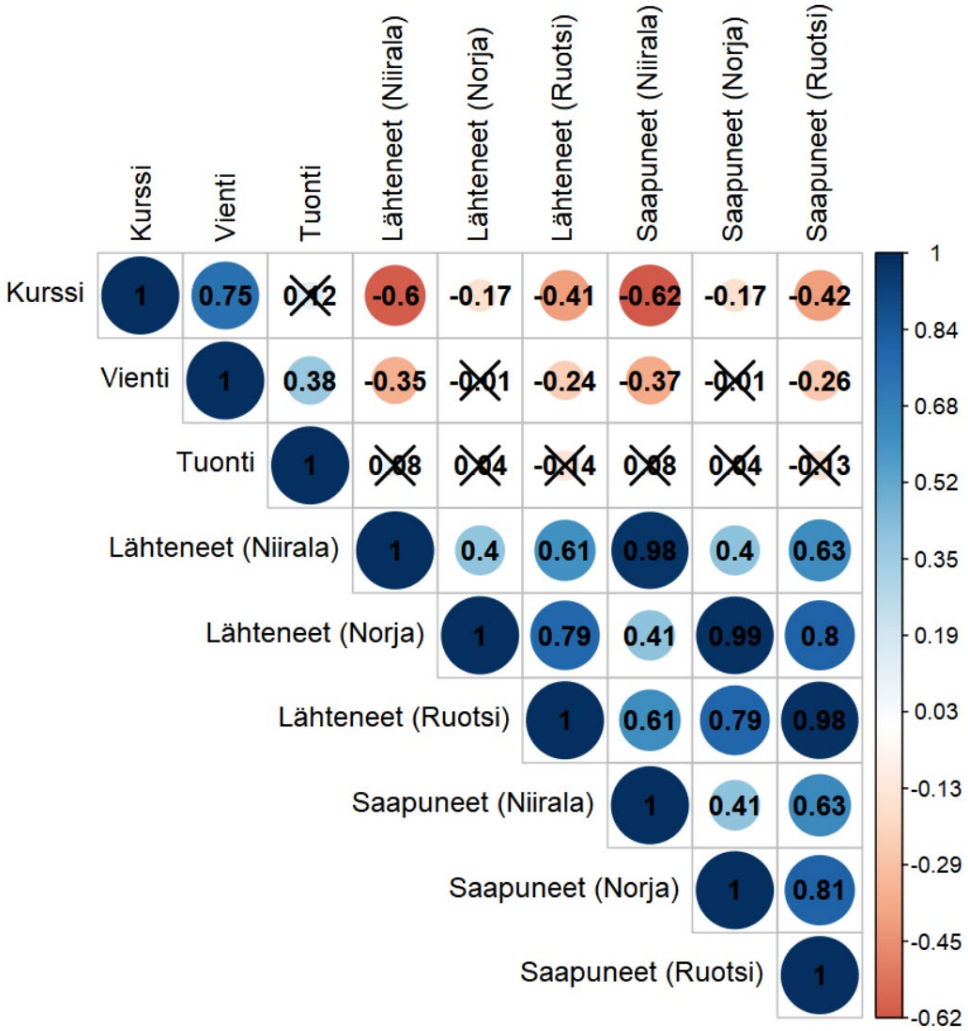
Tämän luvun alussa esitetään korrelaatiomatriisein erilaisten analyysiin sisältyneiden muuttujien keskinäisiä riippuvuussuhteita, eli korrelaatioita. Positiivinen korrelaatio tarkoittaa, että muuttujien arvojen muutokset ajan suhteen noudattelevat tyypillisesti jossain määrin toisiaan. Tässä julkaisussa esitetyt korrelaatiomatriisit sisältävät tarkasteltavien muuttujien nimien leikkauskohdissa korrelaatiokertoimen, sekä havainnollistavan ympyrän, jonka väristä ja koosta voi tulkita korrelaatiokertoimen suuntaa ja voimakkuutta. Ruksilla merkityt korrelaatiokertoimet viestivät siitä, että muuttujien välillä ei ole tilastollisesti merkittävää yhteyttä.

Kuvassa 36 on esitetty joidenkin taloudellisten muuttujien korrelaatiokertoimet rajaliikenteeseen. Tässä korrelaatiomatriisissa lasketaan saman maan osalta sekä lähteneiden että saapuneiden rajanylittäjien määriä kuvaavien muuttujien korrelaatiokertoimet. Luvussa 5.1 esitetyn kaltaisesti samaan maahan kohdistuva lähtevä ja saapuva liikenne ovat tyypillisesti melko lähellä toisiaan, joten näissä tapauksissa korrelaatiokerroin on ymmärrettävästi korkea. Kiinnostavana ilmiöinä esiintyy ruplan kurssin ja tuonnin ei-tilastollisesti merkitsevä riippuvuussuhde, joka eroaa merkittävästi ruplan kurssin ja viennin merkittävästä tilastollisesta yhteydestä 0,75 korrelaatiokertoimella.

Niiralan rajaliikenteeseen suurin vaikutus tarkasteltavista taloudellisista muuttujista näyttää olevan ruplan kurssi -0,6 korrelaatiokertoimella – tämä tarkoittaa sitä, että

ruplan kurssin laskiessa on tarkastelujaksolla rajaliikenteen määrä tyypillisesti noussut. Myös viennin ja rajaliikenteen määrän välillä on havaittavissa tilastollisesti merkitsevä yhteys negatiivisella -0,35 korrelaatiokertoimella, eli viennin pienentyessä on rajaliikenteen määrä tyypillisesti kasvanut. Tuonnin määrällä ei havaittu olevan tilastollisesti merkitsevää yhteyttä rajaliikenteeseen.

Niiralan rajaliikenteellä, sekä Suomen ja Ruotsin, kuin myös Suomen ja Norjan välisellä rajaliikenteellä on positiivinen korrelaatio. Kysymyksessä saattaa olla sesonkihuippujen osuminen tyypillisesti ajanjaksoille, kuten myös mahdollinen samansuuntainen trendi rajanylittäjien kehityksessä, jota ei tosin tässä yhteydessä enää erikseen tarkistettu.

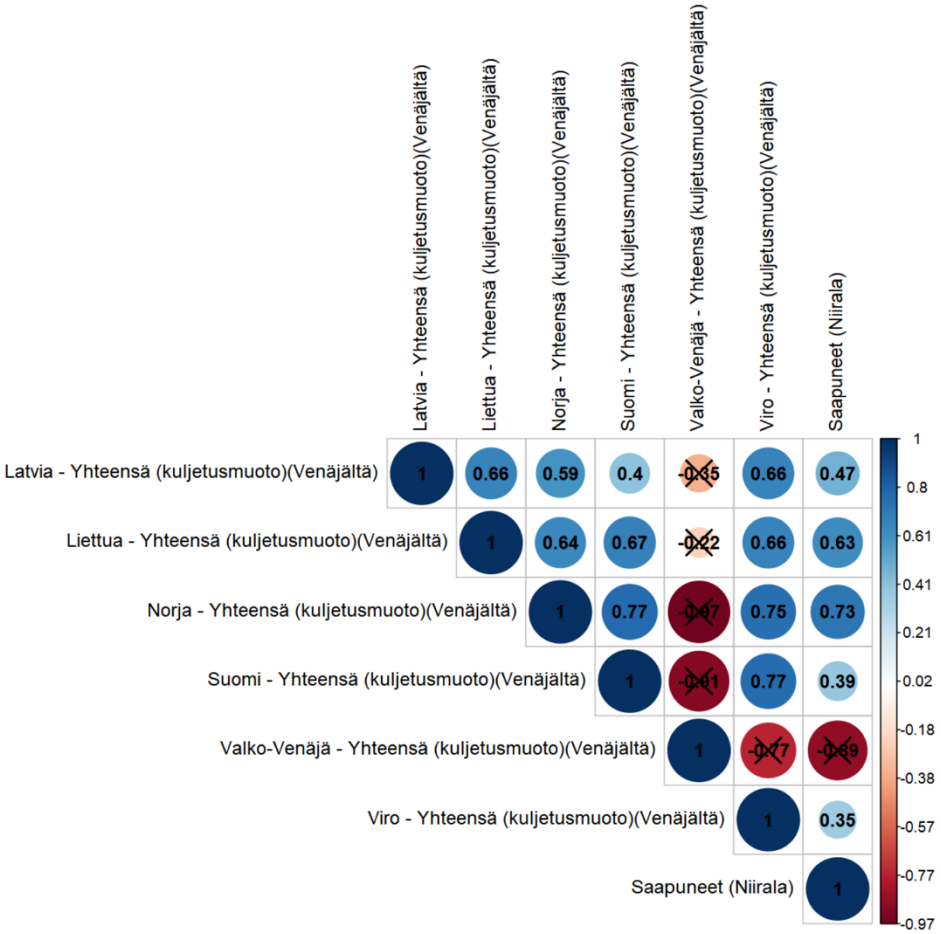


Kuva 36. Taloudellisten muuttujien korrelaatio rajaliikenteeseen

Kuvassa 37 on esitetty Suomea maantieteellisesti läheisten maiden, Suomen sekä Niiralan rajaliikenteen korrelaatiot korrelaatiomatriisissa. Kaikkien maiden rajaliikenteiden kehitysten välillä vallitsee positiivinen korrelaatio tilastollisesti merkityksellisten korrelaatioiden osalta – joitakin negatiivisiakin korrelaatiokertoimia on matriisissa havaittavissa, mutta niiden jäädessä tilastollisesti ei-merkityksellisiksi niiden varaan ei ole mahdollista perustaa uskottavia johtopäätöksiä.

Voimakkain korrelaatio Venäjältä muihin lähimaihin saapuvien rajanylittäjien sekä Niiralan rajanylittäjien määrän välillä oli Venäjältä Norjaan saapuvien sekä Niiralan

rajan kautta Suomeen saapuvien (0,73 korrelaatiokerroin) välillä. Myös Liettuaan Venäjältä saapuvien ja Niiralan rajan yli Suomeen saapuvien tulijoiden välillä oli voimakas tilastollinen yhteys 0,63 korrelaatiokertoimella. Huomionarvoista korrelaatiomatriisissa on se, että Suomeen Venäjältä kokonaisuutena saapuva rajanylittäjien määrä näyttää korreloivan heikommin Niiralan rajanylittäjien määrän kanssa, kuin muihin maihin Venäjältä kohdistuva liikenne. Tätä voidaan pitää hieman yllättävänä tuloksena.



Kuva 37. Maantieteellisesti Suomea lähellä sijaitsevien maiden rajaliikenteen korrelaatio Niiralan rajaliikenteeseen

5.3 Colab-työkalun aikasarjaennusteet

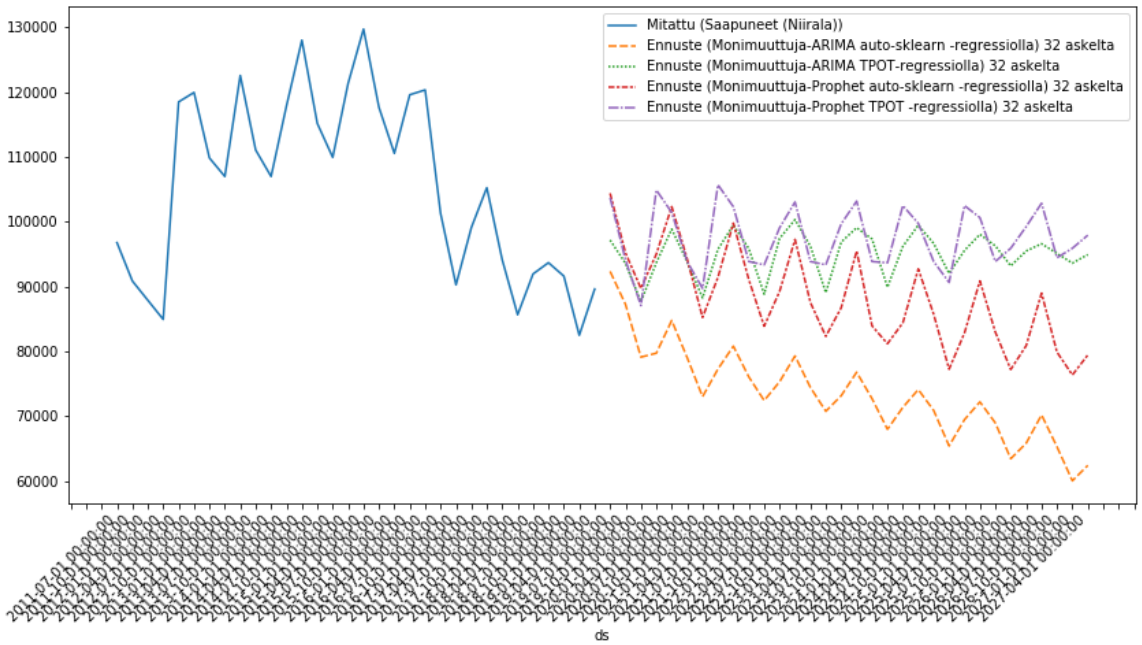
Luvussa 4 kuvatulla Colab-työkalulla toteutettiin Niiralan rajan yli Suomeen saapuvien potentiaalista tulevaisuuden kehitystä kuvastavia ennustemalleja. Mallien onnistuneisuutta on arvioitu tarkemmin luvussa 4.5.2 – kyseisessä luvussa mainitut seikat on hyvä ottaa huomioon tässä luvussa esitettyjä ennusteita arvioitaessa.

On myös hyvä ottaa huomioon, että tulevaisuuteen kohdistuvat ennusteet ovat yleisesti ottaen epävarmoja, sillä niiden tarkkuutta on vertailudatan puuttuessa mahdotonta testata. Lisäksi tulevaisuudessa ennustavissa muuttujissa tapahtuvat mahdolliset trendimuutokset eivät voi ymmärrettävästi sisältyä ennustavista muuttujista tuotettuihin aikasarjaennusteisiin, koska ne eivät ole vielä tiedossa.

Luvussa 4.5.2 esitettyä noudatellen ovat aikasarjaennusteista regressiomallin avulla tuotetut ennusteet pääsääntöisesti luultavasti hieman liian korkeita Niiralasta Suomeen tulijoiden määrän suhteen. Toisaalta mikäli oletetaan luvussa 4.5.2 kuvatun mukaisesti saapuneiden määrään vaikuttavan jokin ulkopuolinen vaikuttaja (tai luultavammin useampia vaikuttajia), joka ei ilmene datasettiin sisällytetyistä muuttujista, saattavat korkeammat ennusteet osua lähemmäs todellisuutta, mikäli kyseisen oletetun vaikuttajan vaikutus esimerkiksi lähitulevaisuudessa heikkenee ja tulijoiden määrän kehitys palaa riippuvaisemmaksi datasettiin sisällytetyistä muuttujista.

Visuaalisesti uskottavimman näköisen ennusteen tuottaa tässä tapauksessa Arimalla tuotetut aikasarjaennusteet, joista on tuotettu regressiomallin avulla ennuste Auto-sklearn -koneoppimistyökalulla. Sama ilmiö oli tämän datasetin ja ennustettavan muuttujan osalta havaittavissa myös luvussa 4.5.2. Ennusteen lähtöpiste on suuruusluokaltaan suunnilleen mitatun datan viimeisen pisteen tasolla, ja trendi näyttäisi ennusteessa jatkuvan samankaltaisena mittausdatan loppuvaiheen trendin kanssa. Tämän ennusteen mukaan Niiralan rajanylityspaikalta Suomeen saapuvien määrä tippuisi noin 60000 tulijaan vuosineljännestä kohti vuoteen 2027 mennessä, mikäli trendi jatkuu samanlaisena kuin mittausdatan loppuvaiheessa. Tämänkään ennustemenetelmän tuottama ennuste ei välttämättä ole uskottava tulos, kun otetaan huomioon koko tarkastelujaksolla tapahtunut tulijoiden nousujohteinen trendi, joka on vasta viimeisimpinä tarkasteluvuosina kääntynyt laskuun.

Työkaluyhdistelmien välillä esiintyy merkittävää vaihtelua kausivaihtelun suuruuden arvion suhteen – visuaalisesti uskottavimman näköinen Auto-sklearn regressio Arima-ennusteista näyttäisi sisältävän pienimmän kausivaihtelukomponentin, kun taas Prophet-aikasarjaennusteista toteutettu Auto-sklearn regressio suurimman.

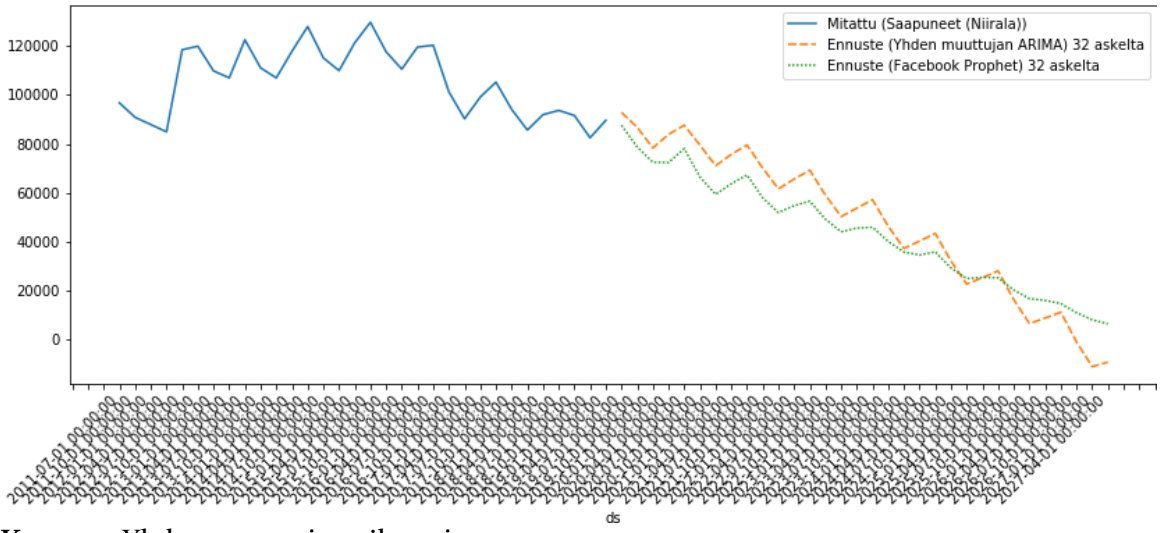


Kuva 38. Kaikkien muuttujien aikasarjaennusteista toteutetut regressiomallit

Yhden muuttujan aikasarjaennusteita tuottavat mallit Facebook Prophet ja Pyramid Arima tuottavat molemmat rajoitteidensa puitteissa uskottavan näköisen tuloksen – ne

ovat tuottaneet toiminnallisuudelleen tyypillisesti datasta trendin sekä kausivaihtelun, ja projisoineet näillä keinoin ennusteen tulevaisuuteen. Mikäli voitaisiin olettaa Niiralan rajanylityspaikalta Suomeen saapuvien määrällisen trendin jatkuvan suoraviivaisesti samankaltaisena kuin se mitatun datan loppuvaiheessa on ollut, olisivat ennusteet varmastikin täysin käyttökelpoisia. Kuten kuitenkin mitatusta datasta voidaan havaita, ei saapuneiden määrä ole tarkastelujaksolla noudattanut mitään selkeää suoraviivaista trendiä, vaan saapuneiden määrässä on tapahtunut trendillisiä muutoksia, joita on vaikeaa tai mahdotonta ennakoida pelkästään ennustettavan muuttujan historiallista kehitystä tarkastelemalla. Näin ollen yhden muuttujan mallit eivät tässä tapauksessa ole luultavasti järkevä lähestymistapa.

Kuten aikaisemminkin on mainittu, saman kaltainen ongelma on myös tässä luvussa esitetyssä moneen muuttujaan perustuvassa regressiomallissa, koska regressiomallin ennuste perustuu yhden muuttujan aikasarjaennusteisiin. Kyseessä voisikin ajatella olevan yleisemmin tulevaisuuden algoritmisen ennustamisen ongelman.



Kuva 39. Yhden muuttujan aikasarjaennusteet

Liite 1. Tietokannat (mukaillen http://uljas.tulli.fi/uljas/ sekä rajapinnan kuvaukset 17.9.2019)

Excel-tietokanta	Selite
/DATABASE/01 ULKOMAANKAUPPATILASTOT/01 CN/ULJAS_CN	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena CN: EU:n yhdistetty tavaranimikkeistö (Combined Nomenclature)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/01 CN/VANHAKK_CN	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena CN: EU:n yhdistetty tavaranimikkeistö (Combined Nomenclature)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/02 SITC/ULJAS_SITC	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena SITC: kansainvälisen kaupan tavaraluokitus (Standard International Trade Classification)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/02 SITC/VANHAKK_SITC	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena SITC: kansainvälisen kaupan tavaraluokitus (Standard International Trade Classification)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/03 CPA/ULJAS_CPA2008	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena CPA: toiminnan lajin mukainen CPA-tuoteluokittelu (Classification of Products by Activities)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/03 CPA/VANHAULJAS_CPA2002	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena CPA: toiminnan lajin mukainen CPA-tuoteluokittelu (Classification of Products by Activities)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/04 TOL/ULJAS_TOL	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena TOL: Euroopan yhteisön toimialaluokitukseen (Nomenclature générale des Activités économiques dans les Communautés européennes, NACE) perustuva kansallinen toimialaluokittelu
/DATABASE/01 ULKOMAANKAUPPATILASTOT/05 BEC/ULJAS_BEC	Ulkomaankauppatilasto kuvaa Suomen ja muiden Euroopan unionin (EU) jäsenvaltioiden ja Suomen ja kolmansien maiden välistä tavarakauppaa eli sisä- ja ulkokauppaa. Luokitteluperusteena BEC: YK:n ylläpitämä SITC-nimikkeistöön perustuva tavaroiden makrotaloudellisen käyttötarkoituksen mukainen luokitus (Classification by Broad Economic Categories)
/DATABASE/01 ULKOMAANKAUPPATILASTOT/06 KAUPPATASE/ULJAS_KAUPPATASE	Kauppatasetilastossa julkaistaan maittaiset kauppatasetiedot. Tilastosta saa myös kauppataseen maan sijaluvun tuonnissa ja viennissä.
/DATABASE/01 ULKOMAANKAUPPATILASTOT/06 KAUPPATASE/ULJAS_KAUPPATASE_VIEW	Kauppatasetilastossa julkaistaan maittaiset kauppatasetiedot. Tilastosta saa myös kauppataseen maan sijaluvun tuonnissa ja viennissä.
/DATABASE/01 ULKOMAANKAUPPATILASTOT/09 INDEKSIT/ULJAS_INDEKSIT	Indeksitilastossa julkaistaan ulkomaankaupan yksikköarvo- ja volyymi-indeksit kuukausittain CPA2008 luokituksen 1-3 nimiketasoilla.
/DATABASE/02 LOGISTIIKKATILASTOT/07 KULJETUSMUOTO/ULJAS_KTAPA	Kuljetustilasto kuvaa Suomen tuonnin ja viennin kuljetusten tonnimääriä kuljetusmuodoittain ja tavararyhmittäin (SITC) sekä alkuperä- ja lähetysmaittain (tuonti) sekä määrämaittain (vienti).
/DATABASE/02 LOGISTIIKKATILASTOT/07 KULJETUSMUOTO/ULJAS_ULKOKONTTI	Kuljetustilasto kuvaa Suomen tuonnin ja viennin kuljetusten tonnimääriä kuljetusmuodoittain ja tavararyhmittäin (SITC) sekä alkuperä- ja lähetysmaittain (tuonti) sekä määrämaittain (vienti).
/DATABASE/02 LOGISTIIKKATILASTOT/08_TRANSITO/ULJAS_TRANSITO	Transitotilastoon sisältyy tiedot maantiekuljetuksista, jotka viedään transitotavarana Suomen läpi itärajan yli tärkeimpien rajanylityspaikkojen (Vaalimaa, Nuijamaa, Niirala ja Imatra) kautta.
/DATABASE/02 LOGISTIIKKATILASTOT/09 RAJALIIKENNE/ULJAS_RAJALIIKENNE	Rajaliikennetilasto kuvaa Suomen rajaliikenteen liikennemääriä liikennevälineittäin ja rajanylityspaikoittain.
/DATABASE/03 VERO- JA KANTOTILASTOT/11 TULLINKANTO/ULJAS_tullinkanto	Tilasto sisältää suoraan valtiolle tuloutettavat verotulot, kantopalkkiot, muut maksut ja sekalaiset tulot, muille viranomaisille tilitettävät maksut, muihin sopimuksiin perustuvan kannon sekä EU:lle tilitettävät tullit ja maksut.

Liite 2. Työkalun ajon aikana syntyvä stats.json -tiedosto


```
{
    "Run type (forcast/test)": "test",
    "Train / test split": 0.9,
    "Metrics": {
        "MvArimaR2": -0.9994214678446747,
        "MvArimaMeanAE": 4554.753526785712,
        "MvArimaMedianAE": 3353.942968749994,
        "MvArimaSQE": 32053683.967577916,
        "TPOTMvArimaR2": -6.060399025734154,
        "TPOTMvArimaMeanAE": 10210.645616213133,
        "TPOTMvArimaMedianAE": 11448.229142857119,
        "TPOTMvArimaSQE": 113188641.16220376,
        "ArimaR2": -4.265483487314049,
        "ArimaMeanAE": 7860.57048964438,
        "ArimaMedianAE": 5369.492699278766,
        "ArimaSQE": 84413489.77852236,
        "ProphetR2": -71.27873358381713,
        "ProphetMeanAE": 33546.10210727454,
        "ProphetMedianAE": 33306.36184490219,
        "ProphetSQE": 1158735024.6718545,
        "MvProphetR2": -28.181463308348874,
        "MvProphetMeanAE": 21276.438169642857,
        "MvProphetMedianAE": 20766.988593749993,
        "MvProphetSQE": 467822026.3412467,
        "TPOTMvProphetR2": -54.4933556116241,
        "TPOTMvProphetMeanAE": 29389.635271905256,
        "TPOTMvProphetMedianAE": 29275.61255342458,
        "TPOTMvProphetSQE": 889640584.3800809,
        "ArimaExoR2": -25.556093131968495,
        "ArimaExoMeanAE": 16359.374029019016,
        "ArimaExoMedianAE": 14841.534994536778,
        "ArimaExoSQE": 425733458.5805357,
        "ArimaExoProphetR2": -501.85120773195615,
        "ArimaExoProphetMeanAE": 86757.99970683848,
        "ArimaExoProphetMedianAE": 88533.01280469653,
        "ArimaExoProphetSQE": 8061448751.338076
    },
    "auto_arima": {
        "Solver": "lbfgs",
        "Max iterations": 50,
        "Seasonality": 4
    },
    "Auto-Sklearn": {
        "Max time": 3600
    },
    "TPOT": {
        "Max time": 60
    },
    "Time series": {
        "Frequency": "QS",
        "Frequency description": "quarter start frequency"
    },
    "Variables": [
        "Saapuneet (Niirala)",
        "Kurssi",
        "Tuonti",
        "Vienti",
        "Kauppatase",
        "L\u00e4hteneet (Ruotsi)",
        "Saapuneet (Ruotsi)",
        "Saapuneet (Norja)",
        "L\u00e4hteneet (Norja)",
        "Suomi - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Ty\u00f6(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Opiskelu(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Auto(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Juna(Ven\u00e4j\u00e4lt\u00e4)",
        "Suomi - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
```

```
    "Suomi - Jalan(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Ty\u00f6(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Opiskelu(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Auto(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Juna(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
    "Viro - Jalan(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Ty\u00f6(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Opiskelu(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Auto(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Juna(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
    "Latvia - Jalan(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Ty\u00f6(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Opiskelu(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Auto(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Juna(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
    "Liettua - Jalan(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Auto(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Juna(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
    "Valko-Ven\u00e4j\u00e4 - Jalan(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Yhteens\u00e4 (syy)(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Liiketoiminta(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Ty\u00f6(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Turismi(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Yksityinen(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Opiskelu(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Pysyv\u00e4 muutto(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Ajoneuvon huoltohenkil\u00f6kunta(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Sotilaat(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Yhteens\u00e4 (kuljetusmuoto)(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Auto(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Ilmailu(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Juna(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Vesi(Ven\u00e4j\u00e4lt\u00e4)",
    "Norja - Jalan(Ven\u00e4j\u00e4lt\u00e4)"
  ]
}
```