VALTIOVARAINMINISTERIÖ

# Koneoppimisen hyödyntäminen KUTI-järjestelmän tiedon laadun parantamisessa sekä automaattiset seurantatyökalut KUTI-järjestelmässä

Kilpailu- ja kuluttajavirasto

VM/2344/02.02.03.09/2018

## Versio 1.0

## 3.7.2019

# Sisällys



## Dokumentin versiohistoria

| Versio | Päiväys | Laatija | Muutoksen kuvaus |
|---|---|---|---|
| 1.0 | 3.7.2019 | Oliver Kostia | |
| | | | |

# 1. Yhteenveto

Tämä dokumentti on uuden toimintamallin tai teknologiaratkaisun toiminnan todentamiseen tähtäävän kokeilun ”Koneoppimisen hyödyntäminen KUTI-järjestelmän tiedon laadun parantamisessa sekä automaattiset seurantatyökalut KUTI-järjestelmässä” loppuraportti.

Kokeilun tavoitteena oli testata koneoppimista Kilpailu- ja kuluttajaviraston tietojärjestelmässä (KUTI). Kokeilussa toteutettiin kaksi uutta toimintoa KUTI-järjestelmään, jotka molemmat hyödyntävät koneoppimista. Uusien toimintojen tavoitteena oli vähentää manuaalista työtä, parantaa tiedon laatua, sekä helpottaa ja nopeuttaa asiantuntijoiden pääsyä oleelliseen tietoon.

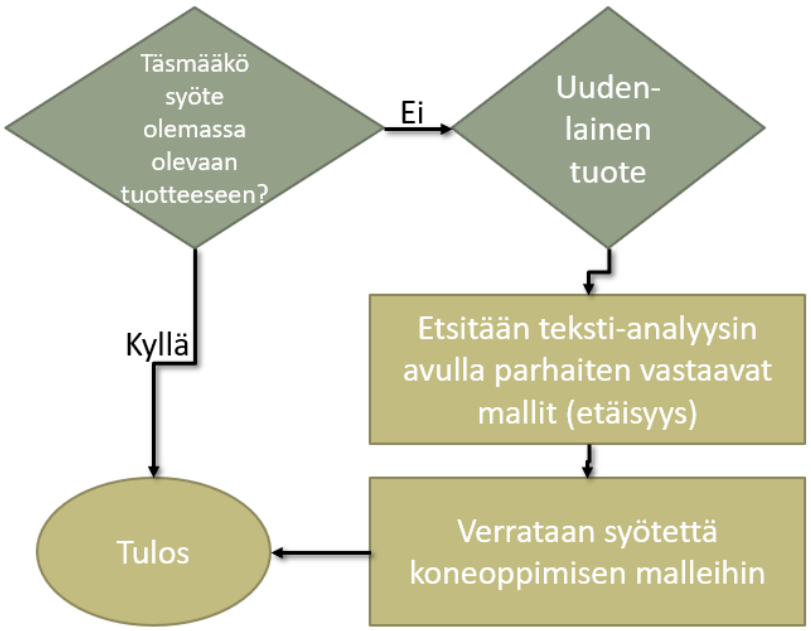
# 2. Kokeilun toteutuminen

## 2.1. Kokeilun tiedot

Kokeilu toteutettiin suunnitelman mukaan käyttämällä Microsoftin .NET ohjelmistokehystä, sekä koneoppimisen toteuttamiseksi avoimen lähdekoodin ML.NET ja Accord.NET -kirjastoja.

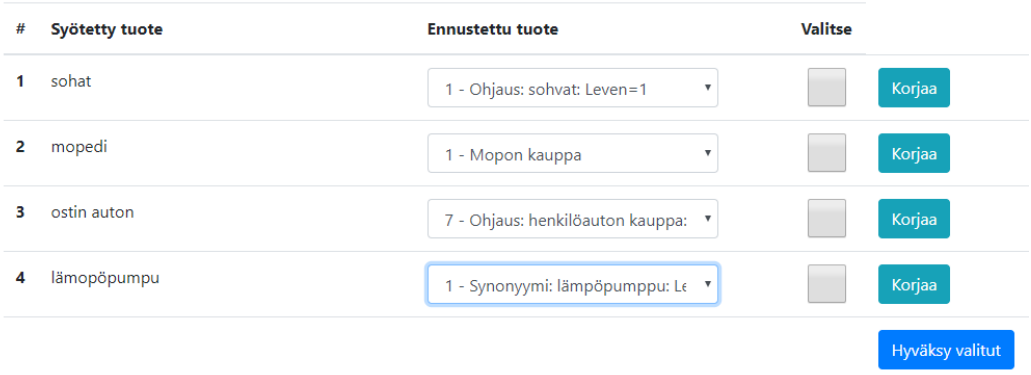
Tiedon laadun parantamiskokeilussa koneoppiva toiminto toteutettiin hyödyntämällä ML.NET kirjaston KMeans-algoritmia, sekä erillistä Levenshteinin etäisyysalgoritmia. Uuden toiminnon tarkoituksena oli löytää järjestelmään tulevalle uudelle sanalle oikea vastine olemassa olevien ns. hyväksyttyjen sanojen joukosta ilman käyttäjän toimenpiteitä. Toiminnossa käytettiin järjestelmästä valmiiksi löytyviä tuotesanoja.

KMeans-algoritmin toiminta perustuu opetusdatan klusterointiin, jolloin samankaltaiset sanat liitetään samaan klusteriin. Ideana oli hyödyntää KMeans-algoritmin luomaa mallia siten, että uudelle tuntemattomalle sanalle etsitään sille lähin klusteri, ja siitä klusterista valittaisiin lähin vastine. Levenshteinin etäisyysalgoritmi toimii siten että se etsii olemassa olevista sanoista sen, joka on lähimpänä uutta annettua sanaa. Lähin sana löydetään laskemalla, kuinka monta muutosta toisen sanan kirjaimiin tarvitaan, että uusi ja vanha sana ovat samanlaiset. Toiminnossa hyödynnettiin molempia algoritmeja.




*Kuva 1 Prosessikuva tiedonlaadun korjaustoiminnosta.*

Nykyisessä versiossa toiminto vastaanottaa sanan, jolloin sille aluksi etsitään täysin samaa vastinetta. Jos vastinetta ei löydy, haetaan tuotteelle vastinetta sekä KMeans-, että Levenshtein-algoritmilla. Nämä tulokset tuodaan käyttäjälle listaan siten, että ”paras” vaihtoehto on listan ensimmäisenä. Tästä listasta käyttäjä voi valita, mihin olemassa olevaan tuotteeseen uusi tuote ohjataan. KMeans-algoritmin opetusmalli päivitetään aina öisin, jolloin toiminto oppii tehdyistä ohjauksista ja tarjoaa parempia vaihtoehtoja uusille tuotteille. Aluksi toiminto vaatii käyttäjän hyväksynnän ohjauksille, mutta jos koneoppiva toiminnon luotettavuus saadaan tarpeeksi korkealle, voitaisiin osa ohjauksista tehdä automaattisesti.



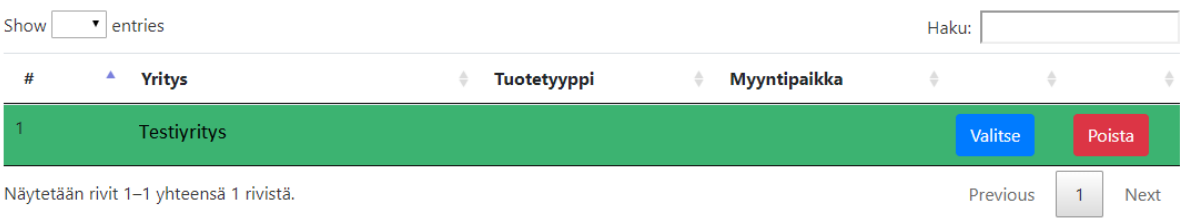

*Kuva 2 Kuva tiedon laadunkorjaustyökalun käyttöliittymästä. (Testiversio)*

Automaattiset seurantatyökalut -toiminnossa koneoppiva toiminto toteutettiin hyödyntämällä Accord.NET -kirjastoa. Toiminnon ideana oli havaita poikkeavuuksia KUTI-järjestelmään saapuvasta datasta (engl. Anomaly detection). Tämän toteuttamisessa hyödynnettiin Accord.NET -kirjastosta löytyvää Support

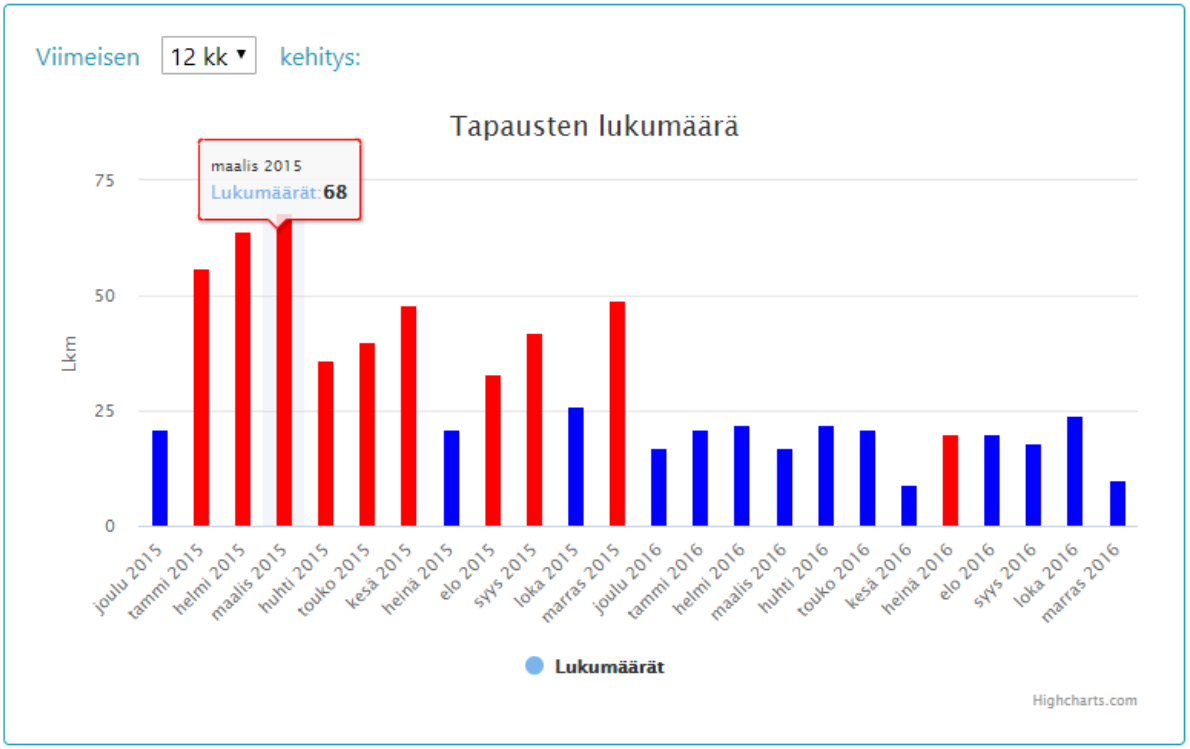
Vector Machinea (SVM, suomeksi Tukivektorikone). Uuden toiminnon tarkoituksena oli löytää järjestelmään tulevasta tiedosta poikkeavuuksia, joiden avulla asiantuntijat pääsisivät käsiksi poikkeukselliseen tietoon nopeasti ilman manuaalisia toimenpiteitä. Sen lisäksi käyttäjille luodaan sovellukseen uusi näkymä, jossa voidaan helposti seurata tiettyjä ajankohtaisia kokonaisuuksia. Näistä annetaan käyttäjälle tieto, jos seuratuissa kokonaisuuksissa on tapahtunut oleellisia muutoksia.


*Kuva 3 Prosessikuva automaattisesta seurannasta.*

Nykyisessä versiossa jokainen käyttäjä voi lisätä listalle haluamansa seurattavat yritykset, joista hän saa yrityskohtaisia lukumäärätietoja. Tämän lisäksi toiminto etsii tulevasta datasta poikkeavuuksia, esimerkiksi tietystä yrityksestä tulevien tapausten lukumääristä. Ideana on lisätä toimintoon mahdollisuus seurata kombinaatioita, jolloin pelkän yrityksen lisäksi voitaisiin esimerkiksi seurata myös kyseisen yrityksen tiettyä tuotetta.




*Kuva 4 Kuvia nykyisestä seurantatyökalun käyttöliittymästä. (Testiversio)*




*Kuva 5 Kuvia nykyisestä seurantatyökalun käyttöliittymästä. (Testiversio)*

## 2.2. Kokeilun rahoitus, kustannukset ja henkilötyöpäivät

Kokeilun suunnitellut (käyttö- ja kirjausoikeuspäätöksen mukaiset) ja toteutuneet kustannukset euroina ovat eriteltyinä omaan ja ostettuun työhön sekä muihin kustannuksiin seuraavat:

| **Kustannus** | **Suunniteltu €** | **Toteutunut €** |
|---|---|---|
| Oma työ (nykyresursseilla tehtävä työ) | 6000 | 4600 |
| Oma työ (kokeiluun erikseen palkattavien resurssien työ) | 0 | 0 |
| Palvelujen ostot | 36 000 | 31960 |
| Muut kustannukset | 0 | 0 |
| **Kokonaiskustannus** | **42000** | **36560** |

Kustannukset eriteltynä rahoituslähteittäin euroina ovat seuraavat:

| **Rahoituslähde** | **Suunniteltu €** | **Toteutunut €** |
|---|---|---|
| 28.70.22 Hallinnon palveluiden digitalisoinnin tuki | 36000 | 31960 |
| 324001 | 6000 | 4600 |
| **Kokonaiskustannus** | **42000** | **36560** |

Kokeilun kustannusten ylitys/alitus johtuu pääosin seuraavista tekijöistä:

- Kustannusten alitus johtuu pääasiassa sovellustoimittajan rajallisista resursseista keväällä

Oman, kokeiluun erikseen palkatun henkilöstön toteutunut kustannus euroina ja henkilötyöpäivinä:

| € | htp |
|---|---|
| 0 | 0 |

## 2.3. Hankintakäytännöt

Hankintaan käytettiin InterTechno Training Oy:n sovelluskehittäjiä, joiden tehtävänä oli toteuttaa kokeilun tekninen kokonaisuus. Kyseinen yritys on toiminut Kuti-järjestelmän kehittäjänä, joten heillä oli erinomainen käsitys siitä, mitä on tarkoituksena toteuttaa.

Kokeiluihin ei tarvittu muita hankintoja.

## 2.4. Riskienhallinta

Yleinen riskitaso kokeilussa oli melko pieni. Isoimmat riskit olivat koneoppimisen hyödyllisyyden todentamisessa sekä kokeilun viivästymisessä. Virastossa ei ole aikaisempaa kokemusta koneoppimisesta, joten hyötyjen realisoituminen nähdään paremmin vasta käytännön kokemusten kautta. Teoriassa ja testien osalta koneoppimisen hyödyt ovat helposti havaittavissa. Aikataulun osalta viivästyminen oli toisena riskinä, joka osaltaan toteutui. Viivästyminen johtui osin toisen kokeilun alkuperäisen suunnitelman muututtua, kun toimintoon suunniteltuja ominaisuuksia haluttiin käyttäjien toiveista lisää. Myös viimeistely ja tuotantoyhdistely veivät odotettua enemmän aikaa.

Kokeilun riskien tilanne kokeilun päättyessä:

| Riski | Lopullinen tila | Toimenpiteet | Toimenpiteiden vaikutus |
|---|---|---|---|
| Koneoppimisen hyöty | avoin | Suunnittelu, ja testaus. | Teorian ja testien puolesta hyödyt on nähtävissä. Käytännön toteutuksesta sen sijaan saadaan tuloksia vasta pidempiaikaisten kokemusten myötä. |
| Kokeilun teknisen toteutuksen viivästyminen | toteutunut | Aikataulun sopiminen sovellustoimittajan kanssa, sekä tapaamisten lisääminen viivästymisen toteutuessa | Hankkeiden viimeistely ja tuotantoyhdistely vei odotettua enemmän aikaa, minkä takia toista kokeilua ei saatu ennen lomia tuotantoon |

## 2.5. Kokeilun tavoitellut hyödyt ja niiden toteutuminen

Kokeilun ensimmäisen osan (koneoppimisen hyödyntäminen tiedon laadun parantamisessa) hyödyt ovat testiversiossa jo osin havaittavissa. Tuotannon kokemuksia saadaan paremmin selville ajan kanssa.

Tavoitteena on:

1. Vähentää korjaustyöhön käytettyä aikaa, jolloin laatu paranee nopeammin ja työaikaa jää enemmän tärkeämpään työhön
2. Laadun paranemisen myötä tiedon avaaminen viraston ulkopuolelle mahdollistetaan

3. Parempi laatu toimii henkilöstön tukena esimerkiksi, päätöksenteossa, valistuksessa ja viestinnässä

Toisen osan (automaattiset seurantatyökalut) hyödyt saadaan paremmin selville tuotantokäytön jälkeen. Testiversiossa toiminto on viimeistelyä vaille valmis, ja teoriassa suunnitellut hyödyt ovat nähtävissä.

Tavoitteena on:

1. Löytää suuresta aineistosta uusia ja piileviä kuluttajaongelmia, jolloin niihin pystytään puuttumaan aikaisemmassa vaiheessa
2. Luoda paremmat työkalut tehtyjen toimenpiteiden seuraamiseen, jolloin saadaan selkeitä mitattavia tuloksia toimenpiteiden vaikuttavuudesta
3. Säästetään tiedon seurantaan ja selaamiseen käytettyä työaikaa

Kuvaa alla olevaan taulukkoon kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaali hakemuksen mukaan ja arvioi sen toteutumista kokeilun jälkeen:

| **Arvio kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaalista** | | |
|---|---|---|
| Tavoiteltava yhteiskunnallinen vaikuttavuus | Hyötyjen realisoituminen hakemuksen mukaan | Arvio hyötyjen realisoitumisen toteutumisesta, jos kokeilussa rakennettu muutos otetaan tuotantoon |
| Mahdollistaisi nopeamman ja tehokkaamman valvonta- ja viestintätyön | Nopeampi reagoiminen kuluttajaongelmiin | 3-24kk riippuen automaattisen seurannan tuloksien hyödyistä, sekä toiminnon käyttöasteesta |
| Viraston valvonnan, vaikuttamisen, valistuksen ja viestinnän vaikuttavuus kuluttaja-asioissa | Tietopyyntöihin ja viestintään saataisiin tarkempaa tietoa. Valvonta- ja vaikuttamistoimenpiteissä asiantuntijoilla olisi parempaa tietoa päätöksenteon tueksi.<br>Neuvontatyössä kuluttajaoikeusneuvojilla on avun antamisen tueksi parempaa tietoa | 4-12kk riippuen kuinka nopeasti laatua saadaan parannettua ja sitä kautta hyödynnettyä |
| Tiedon avaaminen organisaation ulkopuolelle | Tietoa voitaisiin avata kuluttajille, yrityksille, sekä muille viranomaisille | 12+kk riippuen koska tietoa päätetään avata, ja mitä tietoa avataan |
| | | |
| | | |
| | | |

Kuvaa alla olevaan taulukkoon kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaali hakemuksen mukaan ja arvioi sen toteutumista kokeilun jälkeen:

| **Arvio kehitettävän prosessin tuottavuuspotentiaalista** | | |
|---|---|---|
| Taloudelliset hyödyt | Hyötyjen realisoituminen | Arvio hyötyjen realisoitumisen toteutumisesta, jos kokeilussa rakennettu muutos otetaan tuotantoon |
| Henkilöstöresurssien vapautuminen | Työajan säästö automatisoimalla manuaalisia toimenpiteitä | 0 – 3kk riippuen toiminnon tehokkuudesta |

| | | |
|---|---|---|
| Toiminnan tehostuminen | Laadukkaammalla tiedolla päästäisiin parempiin tuloksiin | 3+kk riippuen kuinka nopeasti tiedon laatua saadaan parannettua |
| Piilevät ongelmat | Automaattisella seurannalla päästään käsiksi piilevään tietoon | 3-12kk riippuen automaattisen seurannan tuloksien hyödyllisyydestä tuotantokäytöstä |
| Vaikuttavuuden seuranta | Paremmat työkalut toimenpiteiden vaikuttavuuden seuraamiseksi | 6-12kk riippuen käyttöasteesta, sekä toiminnon tehosta |
| | | |
| | | |

# 3. Kokeilun päättäminen

## 3.1. Kokeilun opit

Kokeilussa opittiin, kuinka koneoppimista hyödyntävä toiminto toteutetaan olemassa olevaan järjestelmään. Itse koneoppimisen osalta todettiin se, että tekstin käsittely suomeksi on vielä melko haastavaa, joten numerodatalla koneoppimisen toteutus olisi ollut helpompaa ja tulokset todennäköisesti vieläkin paremmat.

Algoritmeista opittiin sen verran, että yksi algoritmi ei usein ratkaise koko ongelmaa, vaan toimii osana ongelma ratkaisua. Sen lisäksi todettiin, että algoritmeja voi myös yhdistellä parhaaseen lopputulokseen pääsemiseksi. Algoritmien optimointi on aikaa vievää, joten siihen on hyvä varata aikaa.

Koneoppimisen toteuttaminen ”käsityönä” on iso töistä ja kallista, joten valmiiden palveluiden ostaminen voisi olla halvempaa ja tehokkaampaa. Eli jatkossa täytyy punnita tarkemmin vaihtoehtoja, toteutetaanko räätälöity vaihtoehto, vai etsitäänkö markkinoilta olemassa olevaa palvelua joka ratkaisisi olemassa olevan ongelman.

Jo kyseisen hankkeen aikana koneoppimisen osalta on tullut uutta tekniikkaa ja palveluita, joten se kehittyy selvästi vauhdilla. Nykyään tarjotaan myös pilvipalveluina koneoppimisen toteutukseen palveluita, joilla maaliin pääseminen nopeutuu todennäköisesti huomattavasti. Esimerkiksi on työkaluja, joilla voidaan nopeasti testata, mikä algoritmeista antaa parhaan tuloksen ratkaistavaan ongelmaan nopealla laskennalla, jolloin algoritmin valintaan käytetty aika vähenee huomattavasti.

## 3.2. Kokeilun kokemusten jakaminen

Kokeilusta aiotaan jakaa kokemuksia organisaation sisällä ja sen ulkopuolella. Parempia kokemuksia kokeilun onnistumisesta saadaan ajan kanssa, kun tuotantokokemuksia on saatu enemmän.

Kokemusten jakaminen organisaation ulkopuolelle on osaltaan haastavaa, koska koneoppimisen kokeilun kohteena oli hyvin yksityiskohtainen toiminto viraston omaan tietojärjestelmään. Sen sijaan yleisluontoisia kokemuksia voidaan koneoppimisesta jakaa. Kokemuksia on tarkoitus jakaa VM:n päätöstilaisuudessa sekä erinäisissä verkkopalveluissa.

Organisaation sisällä esitellään hankkeen tuloksia ja arvioidaan, voiko koneoppimista hyödyntää viraston muissa tietojärjestelmissä, tai löytyykö virastolta muita käyttökohteita koneoppimiselle.

## 3.3. Kokeilun hyödyntäminen

Kokeilun tarkoitus oli saada uusia koneoppimista hyödyntäviä toimintoja KUTI-järjestelmän tuotantokäyttöön ja samalla saada kokemuksia koneoppimisen hyödyntämisestä. Tavoitteena oli saada molemmat toiminnot tuotantoon ennen kesälomia 2019, mutta toisen hankkeen osalta viimeistely ja tuotantoyhdistely viivästyi alkusyksyyn 2019.

Tuotantokokemusten läpikäynti tehdään loppuvuodesta 2019, jolloin saadaan konkreettisempaa dataa siitä, onko kokeilun toiminnoista ollut tavoiteltuja hyötyjä, ja tulisiko toimintojen jatkokehitystä harkita.

Kokeilun tuotokset jäävät viraston KUTI-järjestelmään, mutta niiden hyödyntämistä viraston muissa järjestelmissä voidaan harkita laajemman käyttökokemuksen jälkeen.