

Tehisintellekti rakendamise 2. töötuba

14.02.2022



RIIGI INFOSÜSTEEMI AMET



Euroopa Liit
Euroopa
Regionaalarengu Fond



Eesti
tuleviku heaks

TARK



e-RIIK

STACC

Päevakava

09.30–10.00 Kogunemine ja tervituskohv

10.00–11.00 Andmete roll tehisintellekti projektides ning esmased tegevused andmetega. Andmed ennustamiseks

11.00–12.00 Praktiline osa

12.00–12.30 Paus

12.30–13.30 Andmetöötluse olulised aspektid

13.30–14.00 Praktiline osa

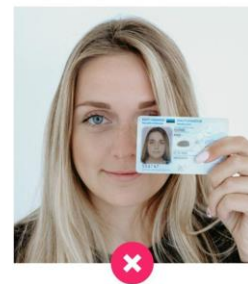
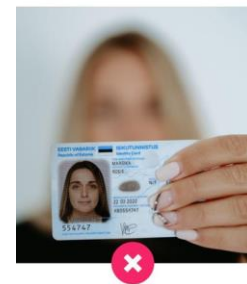
Andmete roll tehisintellekti projektides ning esmased tegevused andmetega. Andmed ennustamiseks

Milline on andmete roll masinõppe ja tehisintellekti projektides?

- Mis on andmed ja nende tähtsus?
- Kuidas andmeid rakendatakse masinõppe ja tehisintellekti mudelites?
- Mõisted: vaatlus/andmepunkt ja tunnus
- Kombinatsioon {01.01.2020, 5, 90, 2010, Tartu} -> 199 500

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tallinn	150 000
...					

Näiteid masinõppe ja tehisintellekti rakendustest Eesti ettevõtete poolt



Veel näiteid masinõppe ja tehisintellekti rakendustest

Eesti ettevõtete poolt

ARABIA
 Muumi Kruus 300ml Little MY and Meadow
 19,99 €

Saadavus:
 E-pood: Saadaval
 Tallinn: Ei ole saadaval
 Tartu: Ei ole saadaval

LISA OSTUKORVI Lisa soovikorvi

Kirjeldus
 Muuminõud on ajatu klassika. Need sobivad nii igapäevaseks kui ka pidulikuks sündmuseks.
 Vitroportselanist valmistatud kruus on igapäevaseks kasutamiseks väga vastupidav.
 Võib kasutada mikrolaineahjus.
 Kuumakindlus kuni 250 °C

Kasutus- ja hooldusjuhised
Lisainfo

SOOVITAME KA NEID TOOTEID



ARABIA
 Taldrik Muumi Misabel
 19cm
 23,99 €



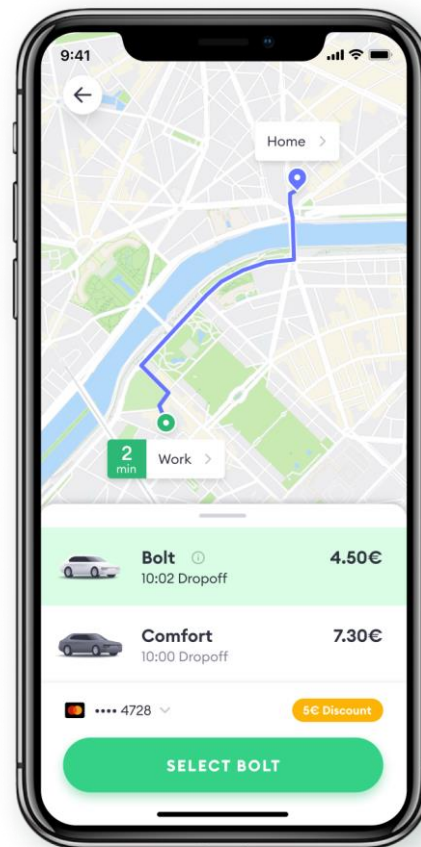
ARABIA
 Muumi Kruus 300ml
 Stinky in Action
 19,99 €



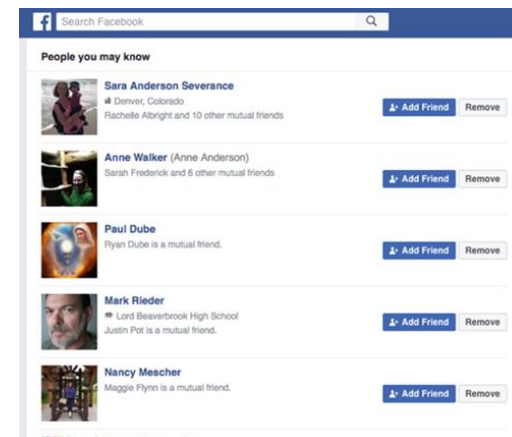
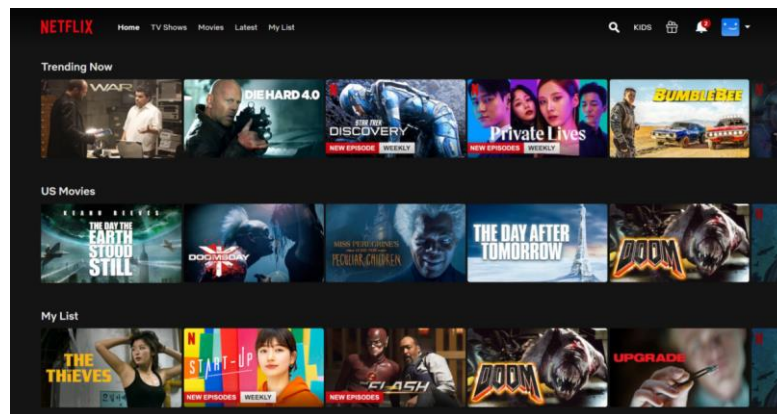
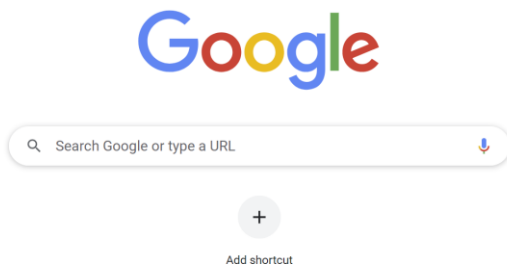
ARABIA
 Muumi Kaus 15cm
 Little MY and Meadow
 23,99 €



ARABIA
 Muumi Taldrik 19cm
 Little MY and Meadow
 23,99 €



Näiteid masinõppe ja tehisintellekti rakendustest maailmas



Näiteid masinõppe ja tehisintellekti rakendustest

- Facebook- Cambridge Analytica skandaal
- NSA programm PRISM
- Hiinas kasutatav näotuvastusprogramm



Masinõppe ja tehisintellekti rakendused avalikus sektoris

- Hinnakokkulepperavimite järelevalve riskimudeli eelanalüüs ja prototüübi loomine.
 - Tellijaks Ravimiamet.
- Ravimi väljamõeldud nimede hindamise tööriist INNA.
 - Tellijaks Ravimiamet.
- Kratt X-tee andmevahetuse anomaaliade tuvastamiseks.
 - Tellijaks Riigi Infosüsteemi Amet.
- Museaalide seisundi ja säilitustähtaegade hindamine - SÄLLI.
 - Tellijaks Muinsuskaitseamet.
- Algamas: Folli-digitaalsete objektide kirjeldamine tehisintellekti abil.
 - Tellijaks Muinsuskaitseamet.



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- Ütlus “Garbage in, garbage out”.
- **Täpsus** (*accuracy*):
 - andmete sisestamise vead,
 - reegel: kui andmed on kogutud inimeste vastustest, siis need on täis vigu,
 - sensorite vead,
 - mõõtmiste ebatäpsused,
 - programmid eksivad,
 - struktureerimata andmete vead.



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- **Terviklikkus** (*completeness*)- vaatluste osakaal, millel kõikide vaadeldavate tunnuste väärtused on olemas ning korrektselt sisestatud.
 - alternatiiv: kui palju on andmeid, mida saab kasutada probleemi lahendamiseks,
 - nõutud andmete hulk sõltub püstitatud probleemist,
 - human level performance test,
 - andmete representatiivsus ning *data balance*,
 - *data drift*.



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- **Vorming** (*format*)- vaatluste vastavus kokkulepitud formaadile.
 - ilupaela pikkus: 1.65 või 165 või 1650?
 - serveri ajatsoon utc 0 või utc +2?
 - süntaks- tunnuste nimed, tabelite nimed, failide nimed.
 - kliendi_id - klient_id

01/11/2021 → 1. november 2021?
 → 11. jaanuar 2021?



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- **Järjepidevus** (*consistency*)- samad tunnused erinevates andmekogudes on identsel kujul.
 - struktureerimata andmete märgendamine.
 - “seal oli ee kass” - ”seal oli ee....kass” - “seal oli kass”.
 - voice activation põhised süsteemid. Näiteks “Hei Siri” on activation key “siri”. Kui palju enne ja pärast sinu juttu lõigata ning millal lõpetada?
 - puuduva väärtuse talletamine - Null, 0, “”, “Null”, NA, “ei oska öelda”.



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- **Duplikaadid** (*duplication*)- vaatluste osalised või täielikud kordused andmekogus.
 - millal on duplikaadid õigustatud?
 - akadeemilises maailmas võib olla halb, väljaspool hea.
 - andmed võivad diskrimineerida.



Andmete kvaliteedist masinõppe ja tehisintellekti vaatenurgast

- Ühenduvus (*integrity*)- olemasolevad seosed andmetes on fikseeritud.
- *concept drift*- kui andmepunktide ja märgendi (tunnus, mida mudeli abil ennustame) seos muutub.
- andmebaasi ülesehitus, dokumentatsioon, välisvõti (*foreign key*) ja primaarvõti (*primary key*).
- andmete koondamine ning ühendamine.



Andmekvaliteedi testimine

RightData
Insight in Sight


DataQ

 talend

QualiDI


xplenty

 QuerySurge

Codoid

iCEDQ

 BiG EVAL


DATAGAPS

Esmased tegevused andmetega

1. Eesmärgi seadmine.
2. Eesmärgi täitmiseks vajalike andmete kombineerimine ja vormindamine.
3. Andmete esmane vaatlus ning profileerimine.
4. Andmete alamosadeks jaotamine – edasiseks analüüsiks vajalike andmete valimine.



Esmased tegevused andmetega

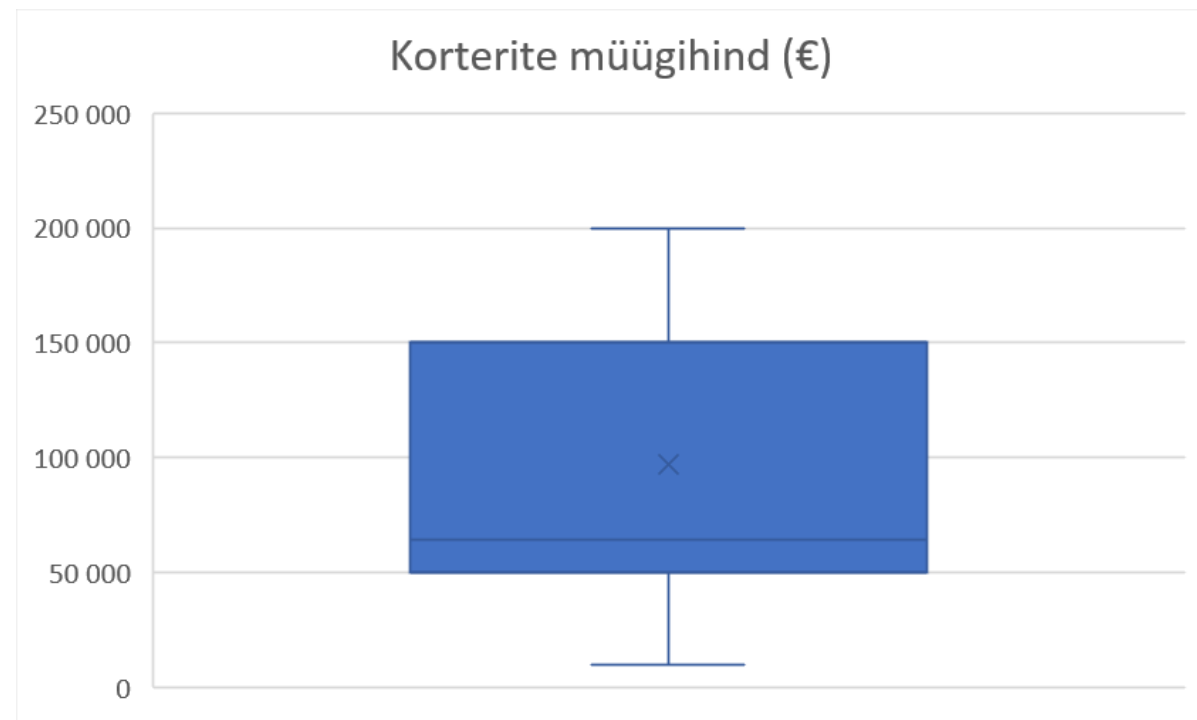
1. Eesmärgi seadmine.
2. Eesmärgi täitmiseks vajalike andmete kombineerimine ja vormindamine.
3. Andmete esmane vaatlus ning profileerimine.
4. Andmete alamosadeks jaotamine – edasiseks analüüsiks vajalike andmete valimine.



ID	Müügikuupäev	Notar	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
84	01.01.2020	Kati Kask	5	90 m ²	2010	Tartu	200 000
789	12.07.2018	Triin Tamm	3	800	2000	Tallinn	50 000
12	3. märts 2019	Martin Mänd	3	65	1989	Pärnu	60 000
109	19.01.2021	Karl Kuusk	3	90	1920	Tllinn	150 000
56	31.12.2021	Helina Haab	4	80	1934	Haapsalu	
...							

Esmased tegevused andmetega

1. Eesmärgi seadmine.
2. Eesmärgi täitmiseks vajalike andmete kombineerimine ja vormindamine.
3. Andmete esmane vaatlus ning profileerimine
4. Andmete alamosadeks jaotamine – edasiseks analüüsiks vajalike andmete valimine.



Esmased tegevused andmetega

1. Eesmärgi seadmine.
2. Eesmärgi täitmiseks vajalike andmete kombineerimine ja vormindamine.
3. Andmete esmane vaatlus ning profileerimine
4. Andmete alamosadeks jaotamine – edasiseks analüüsiks vajalike andmete valimine.

ID	Müügikuupäev	Notar	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
84	01.01.2020	Kati Kask	5	90 m ²	2010	Tartu	200 000
789	12.07.2018	Triin Tamm	3	800	2000	Tallinn	50 000
12	3. märts 2019	Martin Mänd	3	65	1989	Pärnu	60 000
109	19.01.2021	Karl Kuusk	3	90	1920	Tlinn	150 000
56	31.12.2021	Helina Haab	4	80	1934	Haapsalu	
...							



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90 m ²	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
3. märts 2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tlinn	150 000
31.12.2021	4	80	1934	Haapsalu	

Esmased tegevused andmetega

4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete normeerimine ja transformeerimine.

6. Andmete jaotamine treening- ja testandmeteks

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90 m ²	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
3. märts 2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tllinn	150 000
31.12.2021	4	80	1934	Haapsalu	
...					



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90 m ²	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
3. märts 2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tllinn	150 000
...					

Esmased tegevused andmetega

4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete normeerimine ja transformeerimine.

6. Andmete jaotamine treening- ja testandmeteks

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90 m ²	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
3. märts 2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tlinn	150 000
...					



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tlinn	150 000
...					

Esmased tegevused andmetega

4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete jaotamine treening- ja testandmeteks

6. Andmete normeerimine ja transformeerimine.

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
12.07.2018	3	800	2000	Tallinn	50 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tllinn	150 000
...					



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tllinn	150 000
...					

Esmased tegevused andmetega

4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete jaotamine treening- ja testandmeteks

6. Andmete normeerimine ja transformeerimine.

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tlinn	150 000
...					



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tallinn	150 000
...					

Esmased tegevused andmetega

4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete normeerimine ja transformeerimine.

6. Andmete jaotamine treening- ja testandmeteks

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tallinn	150 000
...					



Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)	Hind €/m ²
01.01.2020	5	90	2010	Tartu	200 000	2222,22
03.03.2019	3	65	1989	Pärnu	60 000	923,08
19.01.2021	3	90	1920	Tallinn	150 000	1666,67
...						

Esmased tegevused andmetega

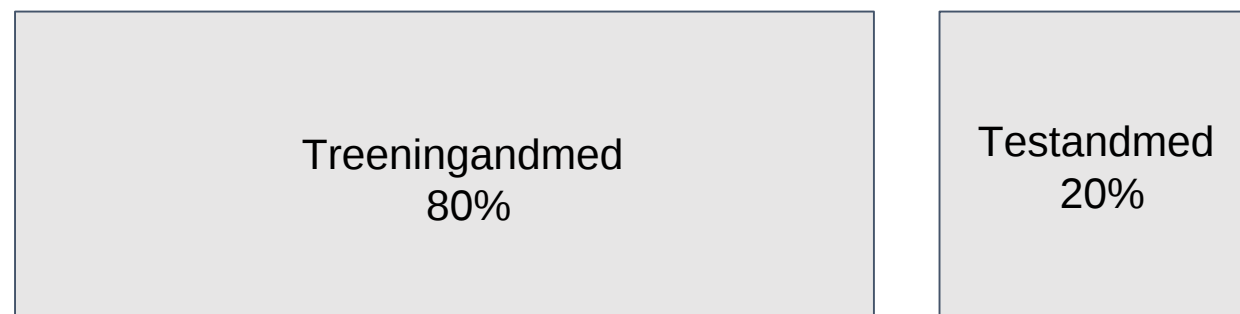
4. Andmepuhastus

- Puuduvate tunnustega vaatluste eemaldamine.
- Valesti vormindatud andmete parandamine või eemaldamine.
- Erindite ja anomaaliate eemaldamine.
- Vigaste andmete parandamine või eemaldamine.

5. Andmete normeerimine ja transformeerimine.

6. Andmete jaotamine treening- ja testandmeteks

Müügikuupäev	Tube	Pindala (m ²)	Ehitatud	Asukoht	Hind (€)
01.01.2020	5	90	2010	Tartu	200 000
03.03.2019	3	65	1989	Pärnu	60 000
19.01.2021	3	90	1920	Tallinn	150 000
...					

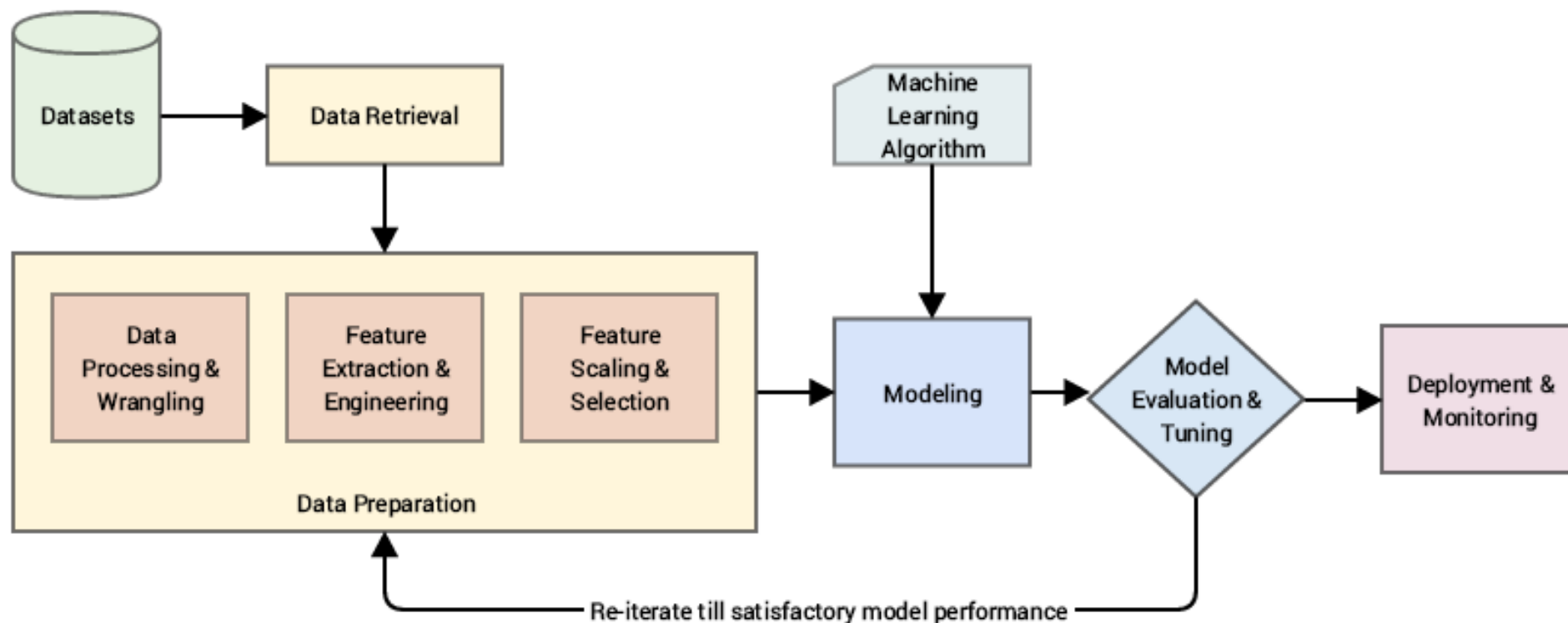


Kas püstitatud probleem on lahendatav?

	area	dry_yield_in_tonnes	crop_name
0	2.89	2.0	Oats
1	10.47	2.0	Oats
2	4.96	2.0	Oats
3	2.98	2.0	Oats
4	14.69	2.0	Oats
5	2.24	2.0	Oats
6	29.42	2.0	Pea
7	9.71	2.0	Winter Rye
8	30.07	2.0	Winter Rye
9	4.80	2.0	Winter Rye
10	34.42	2.0	Winter Rye
11	29.33	2.0	Winter Rye
12	24.14	2.0	Winter Rye
13	3.73	2.0	Winter Rye
14	9.05	2.0	Winter Wheat
15	18.35	2.0	Winter Wheat
16	34.44	2.0	Winter Wheat

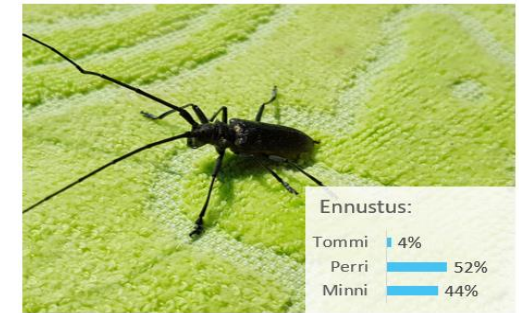
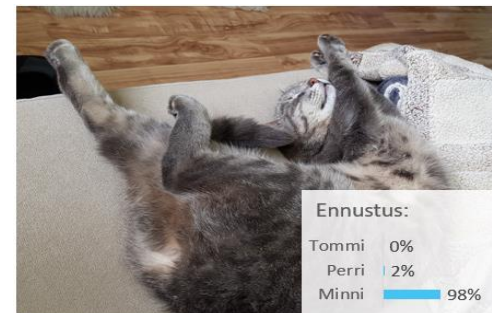
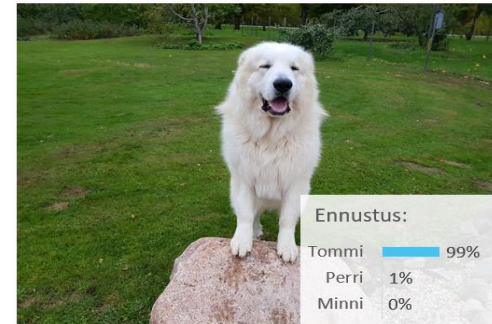
Millega arvestada masinõppe ja tehisintellekti mudelite arenduse puhul?

data drift



Andmelekked masinõppes

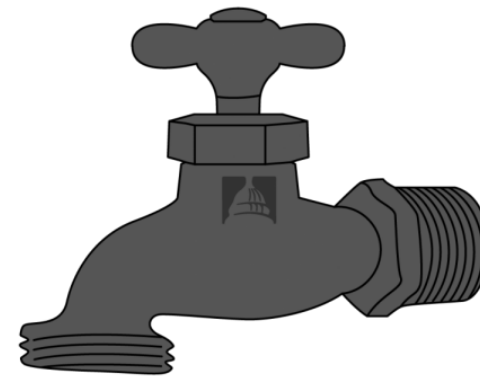
- Kui mudel on väga täpne, on põhjust olla kahtlustav.
 - ülesobitamine (*overfitting*)- mudel õpib treeningandmed ükshaaval pähe ega tee üldistusi, mis võimaldaks uutele näidetele täpselt vastata. Saadakse häid tulemusi treeningandmetel ning hakatakse mudelit liigselt usaldama
- Andmeleke- testandmed lekivad andmetesse, mida kasutatakse mudeli treenimisel.
- Duplikaadid
 - andmestikus olemas
 - tekivad töötlusega
 - andmeteadlase viga



Andmelekked masinõppes

Kuidas vältida andmelekked?

- Korrektne andmete ettevalmistus
- Andmete jagamine treening- ja testandmeteks.
- k-fold cross-validation



011 110
110 01
101 01
100 00
100011
000 011
000 011
110 011
111 001
011 110
1110 011
0101 01110
000 01110100 00100000 0111
0101 0111010001100101 0111

(Tundlike) Andmete lekked

- Pseudonümiseerimine- isikuandmete töötlemine sellisel viisil, et isikuandmeid ei saa enam täiendavat teavet kasutamata seostada konkreetse andmesubjektiga ning täiendavat teavet hoitakse eraldi.
- Tundlikke andmeid tuleks hoida kogus nii vähe kui võimalik, kuid nii palju kui vajalik.
- Andmete hoiustamine.
- Õiguste ja ligipääsude süsteem põhimõttel “need to know”
- Hashimine
- Krüpteerimine

Käesolevas kriminaalasjas süüdistatakse Endel Ooli selles, et tema, olles alkoholihoobes, 20.02.2021 kella 23.15 ajal, viibides Saare maakonnas VVVVVVVVVVVVVVVV talus, ründas noaga toas voodil istunud XX lüües teda sellega ühel korral paremale kõhu

Praktiline töö

<https://tinyurl.com/ai-kratt1>

Paus

12.00–12.30



Riigi Infosüsteemi Amet



TARK



e-RIIK

STACC

Koolituse korraldas Majandus- ja Kommunikatsiooniministeerium
Euroopa Liidu struktuuritoetuse toetuskeemist "Infoühiskonna
teadlikkuse tõstmine", mida rahastab Euroopa Regionaalarengu
Fond.

Koolituse aitas läbi viia Conference Expert / Reisiekspert



RIIGI INFOSÜSTEEMI AMET



Euroopa Liit
Euroopa
Regionaalarengu Fond



Eesti
tuleviku heaks

TARK



e-RIIK

STACC