Anvendelse af AI-analysemetoder til at belyse indhold og datakvalitet i centrale sundhedsregistre

E-sundhedsobservatoriet d. 12. oktober 2023

Henrik Jensen
Kristian Holt Nielsen
Henrik Mulvad Hansen

SUNDHEDSDATA-STYRELSEN

Indhold

- **Baggrund og succeskriterier**
- **Erfaringer og resultater med at anvende AI til datakvalitetsundersøgelser**
  - Case 1: Validering af cancerdiagnoser i Landspatientregistret (LPR)
  - Case 2: Identificere unormale udsving i indberetningsfrekvenser i LPR
- **Perspektiver og videre anvendelse**
- **Spørgsmål**

Baggrund og succeskriterier

Baggrund og formål for projektet

- Voksende volumen, diversitet, og kompleksitet af data som indrapporteres til SDS

  +

- Stigende efterspørgsel efter Sundhedsdatastyrelsens sundhedsdata/sundhedsregistre

  +

- Initiativer igangsat mhp at sikre at sundhedsdata bliver (lettere) tilgængelige for såvel primær som sekundær anvendelse (f.eks. EHDS-forordningen)

  ⬇

- Stigende efterspørgsel på oplysninger om data (tilgængelighed, indhold, datakvalitet, etc)

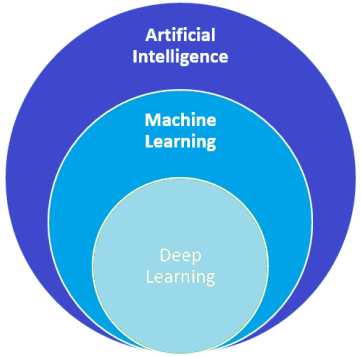
  +

- Behov for at udvikle og effektivisere eksisterende monitoreringer/analyser fra de centrale sundhedsregistre

- *Projektet afprøver hvordan AI baserede metoder kan identificere mønstre og ændringer i indberetninger til SDS sundhedsregistre og dermed*

- *Give overblik over de informationer der findes i sundhedsregistrene, herunder hvilke fejl og mangler der er i indberetningerne*

- *Give mulighed for at reagere proaktivt på fejl/mangler i indberetninger mhp at forbedre kvaliteten af oplysningerne i de centrale sundhedsregistre*



Succeskriterier for AI datakvalitets pilotprojekt

Forbedre datakvalitet

- Identificere ændringer i indberetningsmønstre og indhold
- Afdække fejl og mangler i sundhedsdata
- Få overblik over datakvalitet på tværs af datakilder

AI kompetenceudvikling

- Opbygge kompetencer med AI metoder i SDS
- Sidemandsoplæring og undgå ”black box”
- Udvikling af metoder som kan benyttes på tværs af registre

Erfaringer og resultater med at anvende AI til datakvalitetsundersøgelser

Case 1: Validering af cancer diagnoser i LPR

Problemstilling

Diagnoser er et helt centralt element i indberetningerne til LPR!
Indberetninger med manglende eller ikke-opdaterede aktionsdiagnoser.
Svært at finde ud af hvor omfangsrigt problemet er

Mål

Identifikation af mangler i diagnose indberetning

Data

Der tages udgangspunkt i et udvalgt område – *cancer*.
Metoden kan afprøves på andre områder i senere faser.

Analyser

Klassifikationsalgoritmer, prædiktive algoritmer,
clustering og outlier detection.

Mål og idéer på sigt

Når der indberettes nye oplysninger til LPR,
giver algoritmen en sandsynlighed for at aktionsdiagnosen er korrekt,
baseret på diverse inputvariable.

Ja! AI kan hjælpe med at forbedre datakvaliteten i LPR

Algoritmen identificerer manglende cancerdiagnoser

- Der er i alt fundet **155 manglende cancerdiagnoser** ud af et testsæt på 5000 patienter (ca. 3 %).
- Resultater er valideret ved at sammenligne med cancerdiagnoser fra cancerregisteret.
- Et udsnit af resultaterne er manuelt valideret af SDS cancereksperter.

Algoritmen finder relevante features

- Tre forskellige Machine Learning algoritmer er blevet trænet og testet.
- XGBoost algoritmen er bedst med en **nøjagtighed på 0,92.**
- Algoritmen finder relevante features (explainable AI)

Algoritmens output skal justeres til behovet

- Formålet med projektet var at undersøge om det var muligt at identificere manglende cancerdiagnoser, hvilket er påvist.
- Algoritmen skal justeres og trænes på mere data for at sikre robusthed i forhold til idriftsættelse.

Datastruktur på patientniveau

Endelig datastruktur til algoritmetræning

- 20.000 rækker (patienter) - som har haft en ”DZ031 – Observation pga. mistanke om kræft” i løbet af 2021
- ~2000 kolonner fra kontakter, procedurer og kræftpakkeforløbsmarkører

Person-ID	Alder	Antal kontakter med hovedspeciale Klinisk Onkologi	KPF forløbsmarkør: tilbud om initial behandling	…	…	Har haft cancerdiagnose i LPR3	Algoritme cancerdiagnose sandsynlighed	Algoritme cancerdiagnose
1	71	10	1			0	0,90	**1**
2	25	0	0			0	0,20	**0**
3	76	3	0			0	0,60	**1**
…	…	…		…	…	..	…	…

Sammenligning af nøjagtighed for tre algoritmer

- Sammenligning af nøjagtigheden af prædiktioner for tre forskellige **Machine Learning klassifikationsmodeller**
  - Modellerne har fået samme data som input og
  - Skal prædiktere om en patient bør have en cancerdiagnose eller ej.

Algoritmer	Beskrivelse	Nøjagtighed
Logistic Regression	God til binære problemstillinger, da den tvinger værdier mellem 0 og 1, samt god til at visualisere og analysere feature importance. Go to model.	0,87
K-Nearest Neighbor Clustering	Også en god klassifikationsmodel. Den sammenligner et nyt datapunk med de "nærmeste naboer" den har set under træning.	0,84
XGBoost	State of the art machine learning algoritme til bl.a. klassifikationsproblemer. Treebased boosting model, som bruges på tværs af mange industrier og problemstillinger.	0,92

XGBoost har højest nøjagtighed

Begrebsforklaring

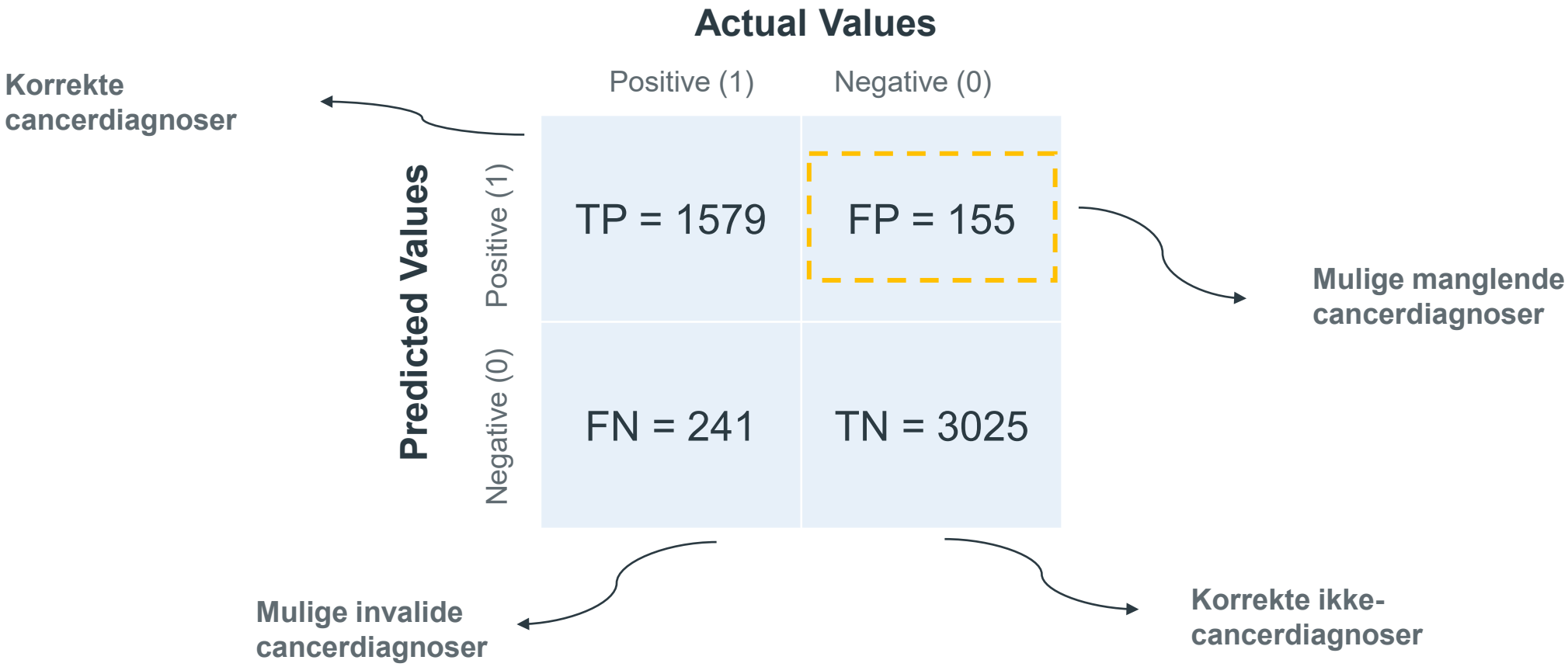
Nøjagtighed

- Er et udtryk for hvor nøjagtig eller pålidelig algoritmen er
- Hvis alle patienters cancerdiagnose er forudsagt korrekt vil nøjagtigheden være 1, hvorimod hvis alle patienters cancerdiagnose er forudsagt forkert vil nøjagtigheden være 0.

$$\text{Nøjagtighed} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{antal korrekt forudsagt}}{\text{total antal}}$$

Algoritmen er bedst, jo nærmere nøjagtigheden er på 1.

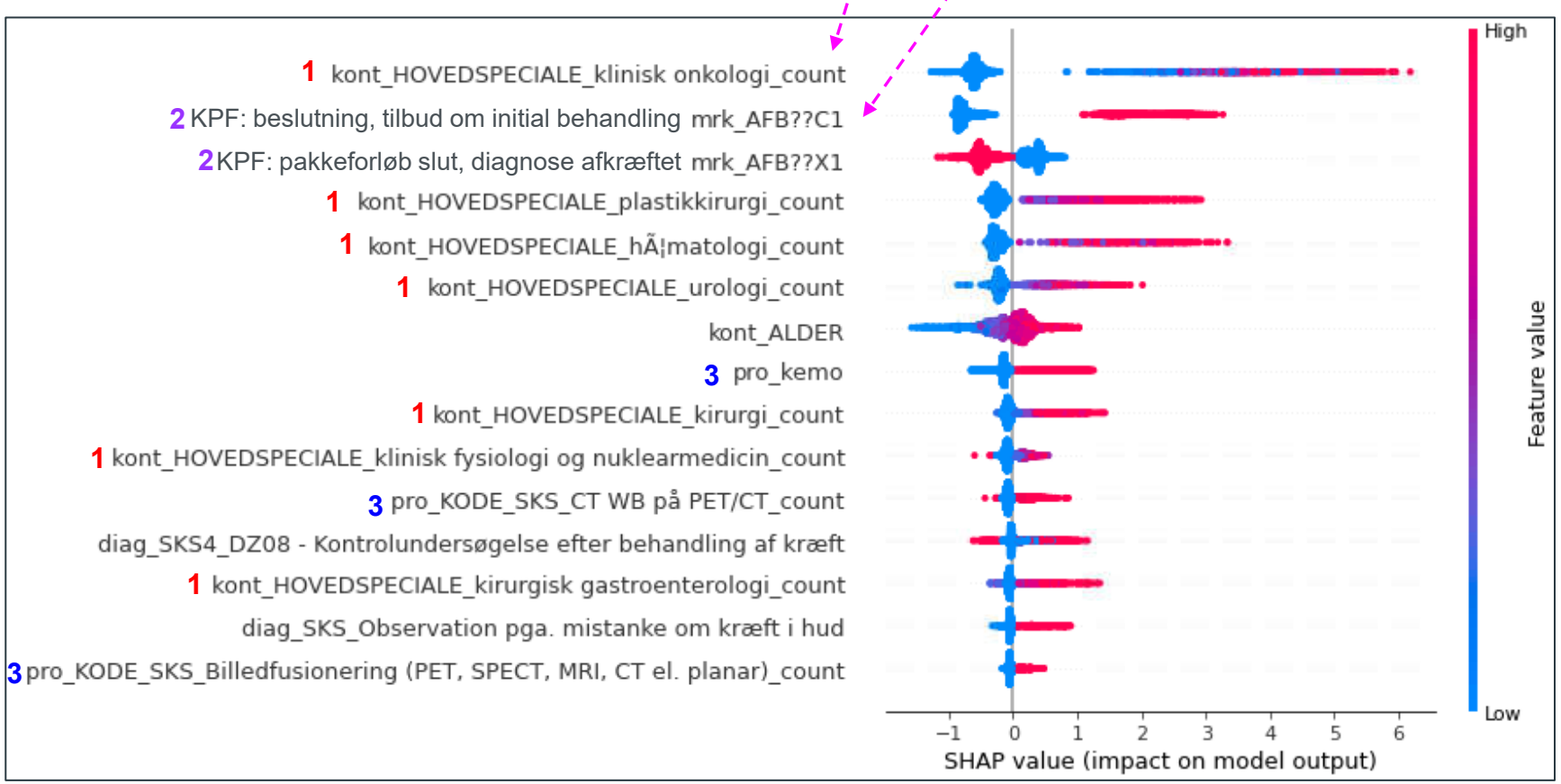
Algoritmens klassifikationer på testsæt af 5.000 patienter



Forklaring af AI modellens (XGBoost) egenskaber (feature importance / Shap værdier)

Count ~ tælle variable (antal)

Øvrige ~ binære variabel (0 eller 1)



Figur forklaring

- Top 15 features (input variable/indberetninger)
- Sorteret efter påvirkningen af hver feature på algoritmens prædiktion af cancer (første række har størst påvirkning).
- En prik på grafen repræsenterer algoritmens prædiktion af én patient i testsættet
    - Farven repræsenterer featurens værdi (høj/lav)
    - x-aksen repræsenterer *SHAP-værdien*/påvirkning af denne feature på algoritmens output
        - Høj værdi ~ høj sandsynlighed for cancerdiagnose
        - Lav værdi ~ lav sandsynlighed for cancerdiagnose

Tolkning

Størst påvirkning på algoritmens forudsigelser af en cancerdiagnose ses for:

1. Antal besøg hos specifikke onkologiske specialer (klinisk onkologi, plastikkirurgi, hæmatologi mm.),
2. Kræftpakkeforløbsmarkører (AFB??C1, AFB??X1)
3. Procedurer (pro_kemo, PET/CT, mm.)

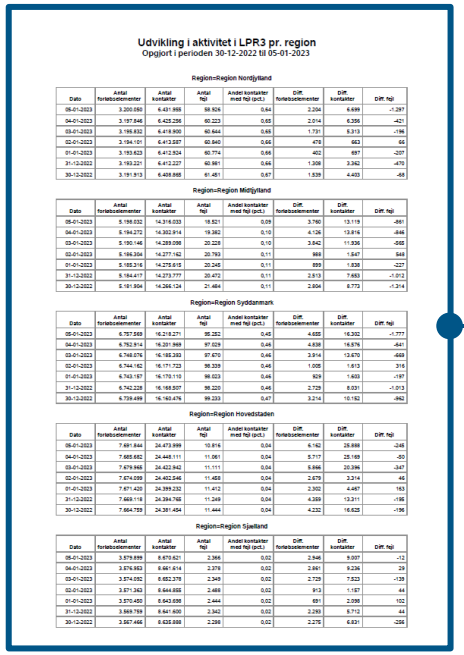
Erfaringer og resultater med at anvende AI til datakvalitetsundersøgelser

Case 2: Unormale udsving i indberetningsfrekvenser i LPR

Identificering af unormale udsving i LPR i dag

Statusmail om LPR

Udarbejdes hver nat og evalueres hver morgen



Udvikling i aktivitet i LPR3 pr. region
Opgjort i perioden 30-12-2022 til 05-01-2023

Region=Region Nordjylland

Dato	Antal forløbselementer	Antal kontakter	Antal hpl	Andel kontakter med hpl (pct.)	Diff. forløbselementer	Diff. kontakter	Diff. hpl
05-01-2023	3.200.050	6.431.988	58.926	0,64	2.204	6.699	-1.297
04-01-2023	3.197.846	6.425.296	60.223	0,63	2.014	6.356	-421
03-01-2023	3.195.832	6.418.900	60.644	0,65	1.731	5.313	-196
02-01-2023	3.194.101	6.413.587	60.840	0,66	478	663	66
01-01-2023	3.193.623	6.412.924	60.774	0,66	402	697	-207
31-12-2022	3.193.221	6.412.227	60.981	0,66	1.308	3.362	-470
30-12-2022	3.191.913	6.408.865	61.451	0,67	1.539	4.403	-68

Region=Region Midtjylland

Dato	Antal forløbselementer	Antal kontakter	Antal hpl	Andel kontakter med hpl (pct.)	Diff. forløbselementer	Diff. kontakter	Diff. hpl
05-01-2023	5.196.032	14.316.033	18.521	0,09	3.760	13.119	-861
04-01-2023	5.194.272	14.302.914	19.382	0,10	4.126	13.914	-846
03-01-2023	5.190.146	14.289.098	20.228	0,10	3.942	11.936	-668
02-01-2023	5.186.304	14.277.162	20.793	0,11	988	1.647	548
01-01-2023	5.185.316	14.275.615	20.245	0,11	899	1.838	-227
31-12-2022	5.184.417	14.273.777	20.472	0,11	2.813	7.653	-1.012
30-12-2022	5.181.904	14.266.124	21.484	0,11	2.804	8.773	-1.314

Region=Region Syddanmark

Dato	Antal forløbselementer	Antal kontakter	Antal hpl	Andel kontakter med hpl (pct.)	Diff. forløbselementer	Diff. kontakter	Diff. hpl
05-01-2023	6.757.569	16.218.271	95.282	0,45	4.655	16.302	-1.777
04-01-2023	6.752.914	16.201.969	97.029	0,46	4.838	16.576	-641
03-01-2023	6.748.076	16.185.393	97.670	0,46	3.916	13.670	-459
02-01-2023	6.744.162	16.171.723	98.339	0,46	1.008	1.613	316
01-01-2023	6.743.187	16.170.110	98.023	0,46	929	1.603	-197
31-12-2022	6.742.228	16.168.507	98.220	0,46	2.729	8.031	-1.013
30-12-2022	6.739.499	16.160.476	99.233	0,47	3.214	10.152	-862

Region=Region Hovedstaden

Dato	Antal forløbselementer	Antal kontakter	Antal hpl	Andel kontakter med hpl (pct.)	Diff. forløbselementer	Diff. kontakter	Diff. hpl
05-01-2023	7.691.844	24.473.999	10.816	0,04	6.162	25.888	-248
04-01-2023	7.685.682	24.448.111	11.061	0,04	5.717	25.169	-50
03-01-2023	7.679.965	24.422.942	11.111	0,04	5.896	20.396	-347
02-01-2023	7.674.069	24.402.546	11.458	0,04	2.679	3.314	46
01-01-2023	7.671.420	24.399.232	11.412	0,04	2.302	4.467	163
31-12-2022	7.669.118	24.394.765	11.249	0,04	4.389	13.311	-196
30-12-2022	7.664.729	24.381.454	11.444	0,04	4.232	16.625	-196

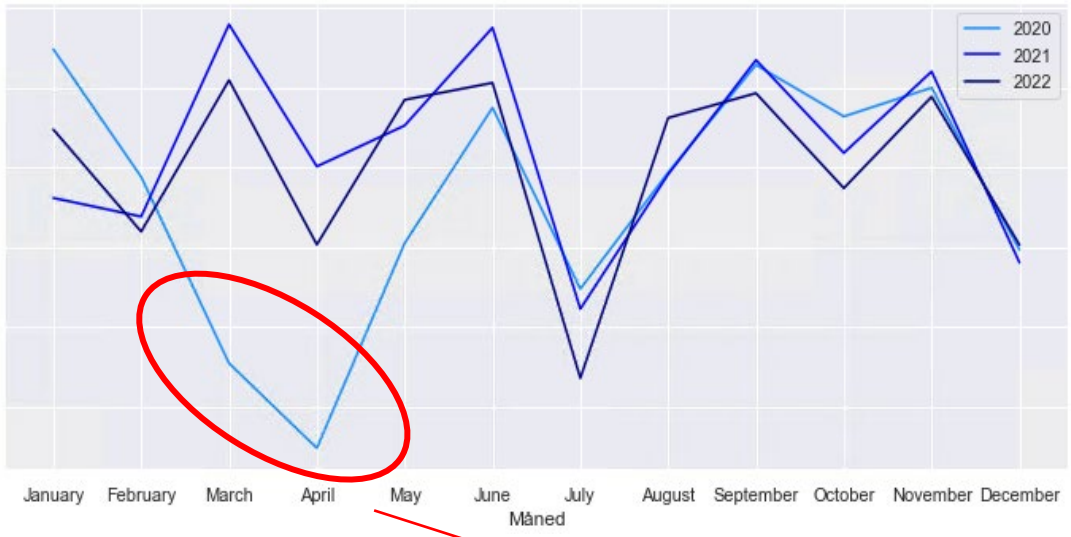
Region=Region Sjælland

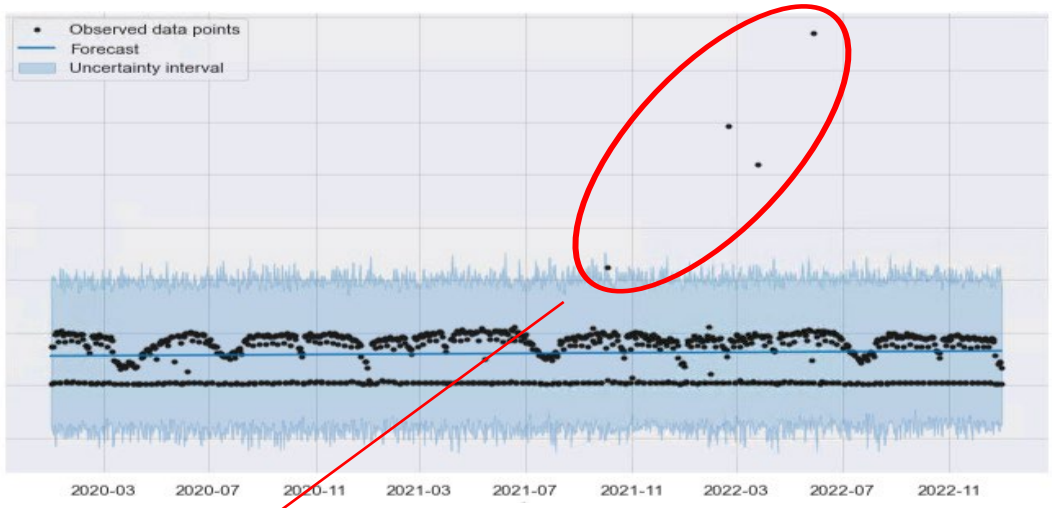
Dato	Antal forløbselementer	Antal kontakter	Antal hpl	Andel kontakter med hpl (pct.)	Diff. forløbselementer	Diff. kontakter	Diff. hpl
05-01-2023	3.579.899	8.670.621	2.366	0,02	2.946	9.007	-12
04-01-2023	3.576.953	8.661.614	2.378	0,02	2.891	9.236	29
03-01-2023	3.574.062	8.652.378	2.349	0,02	2.729	7.523	-139
02-01-2023	3.571.343	8.644.855	2.488	0,02	913	1.187	44
01-01-2023	3.570.430	8.643.668	2.444	0,02	691	2.098	102
31-12-2022	3.569.759	8.641.600	2.342	0,02	2.293	5.712	44
30-12-2022	3.567.466	8.635.888	2.298	0,02	2.275	6.831	-256

Problemstillinger for områdeansvarlig

- Udfordrende at danne overblik over alle indberetninger/områder på samme tid.
- Tidskrævende og en del manuelt arbejde i at undersøge udsving.
- I dag er ”alarmer” regelbaseret. Svært at lave regler for fejl som ikke er sket eller opdaget tidligere.
- Udarbejdes forskellige løsninger på tværs af registre til at overvåge datakvaliteten.
- Flere områder som der ikke er kapacitet til at monitorere i dag.

Identificering af unormale udsving i antal indberetninger over tid





Eksempler på unormale udsving AI skal identificere

Ja! AI kan hjælpe med at forbedre datakvaliteten af LPR

AI identificerer unormale udsving

- AI detekterer de rigtige unormale udsving.
- Vi kan med AI finde andre og mere præcise unormale udsving end med traditionelle metoder
  - Fx identificering af *trend changepoints*
  - Undgår begrænsninger ved traditionelle metoder som fx ved sæsonvariation.

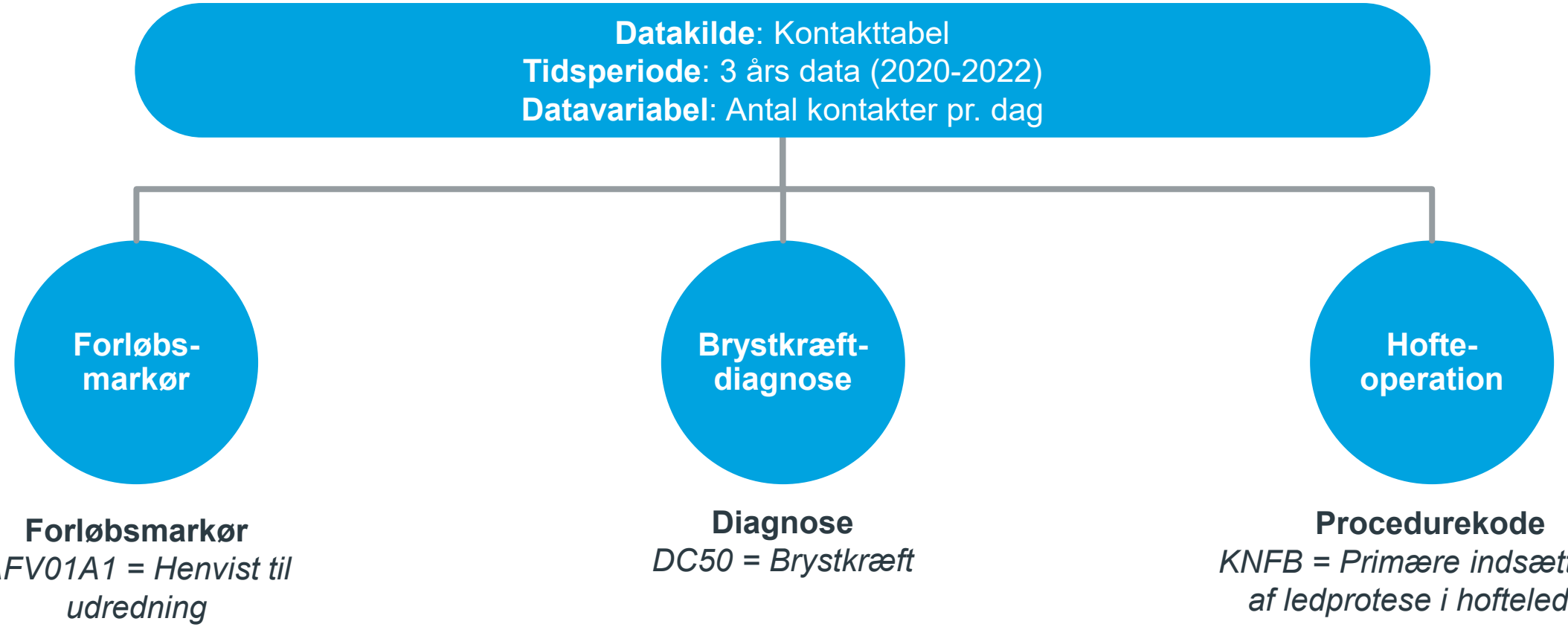
Supervised learning metoder giver bedste resultat

- Både supervised og unsupervised algoritmer er blevet afprøvet.
- Supervised algoritme *(Prophet)* gav bedst performance på tværs af tre use cases.
- Generelt har data stor sæsonvariation, hvilket supervised algoritmer er bedre til at håndtere.

Løsningen er generisk og skalerbar

- Der er udviklet én løsning som kan anvendes på tværs af tre helt forskellige dataområder/use cases.
- Metoden er generisk og kan skaleres til flere dataområder.
- Usikkerhedsinterval kan justeres på tværs af forskellige use cases.

Datakilder for de 3 use cases



Afprøvning af forskellige algoritmer

Machine Learning algoritmer til outlier/anomali detektion

- *Forskellige algoritmer detekterer forskellige typer anomalier.*
- *De algoritmer der performer er beskrevet i nedenstående tabel.*
- *Simplere metoder er også blevet afprøvet, som bl.a. konfidensinterval, moving average, mm.*
  - *Disse er blevet fravalgt grundet dårligere egenskaber til sæsonvariation eller behov for meget manuel tuning.*
- *Vi ønsker at algoritmen skal være generisk og skalerbar.*
- *Alle modeller har fået samme data som input og skal identificere unormale udsving i antal indberetninger over tid.*

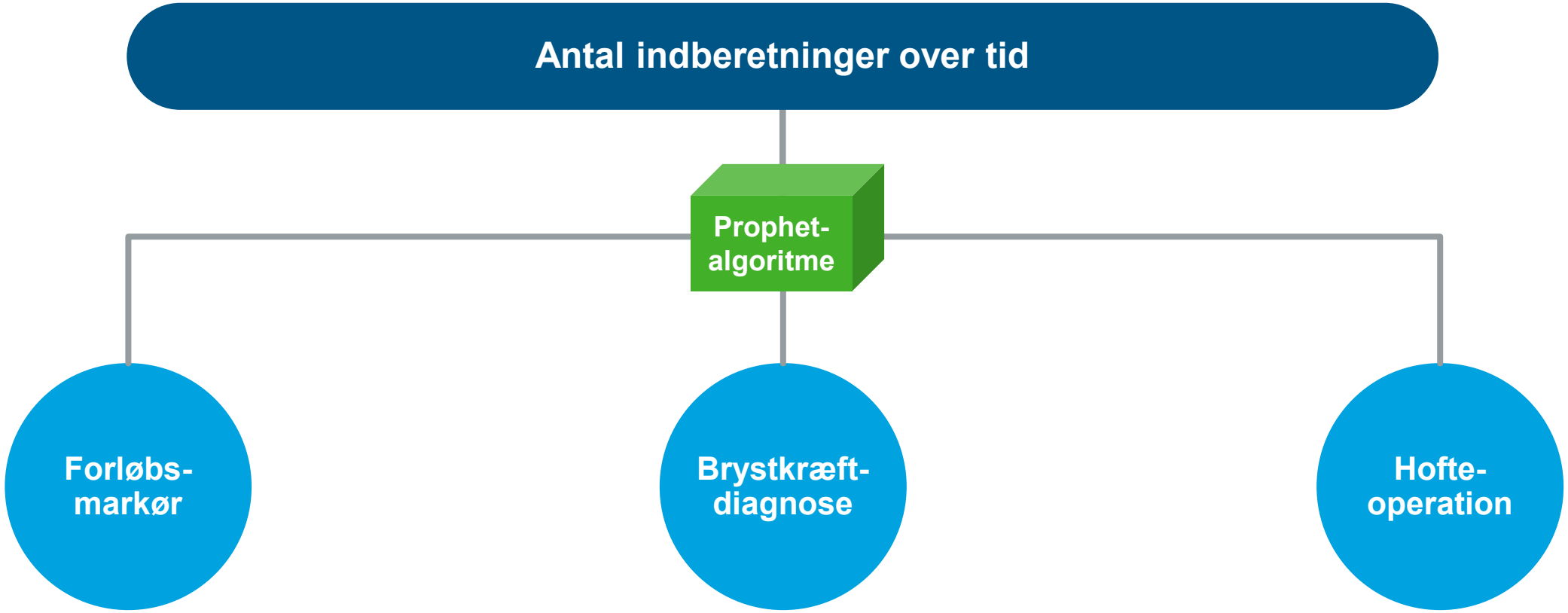
Algoritme	Metode	Beskrivelse	Observationer
Isolation Forest	Unsupervised learning	Isolation forest er en træbaseret algoritme, som i stedet for at kigge efter normale punkter, isolerer de unormale fra de normale punkter ved at fokusere på de anderledes egenskaber.	**Fordele** God til at detektere ekstreme peaks / outliers. Mulighed for at justere hvor stor en procentdel af data der forventes af være outliers (contimination parameter). **Ulemper** Svært ved at håndtere store sæsonvariationer. Identificerer peaks i fx sommermånederne som anomalier.
Prophet	Supervised learning	State-of-the-art algoritme til tidsserieprædiktion udviklet af Facebook's data science team. Anomalier detekteres ved at identificere forskelle i faktiske og prædikterede værdier.	**Fordele** God til at håndtere sæsonvariationer og anomalier. Giver en bedre forståelse for anomali detektion. Mulighed for at justere usikkerhedsinterval. **Ulemper** Virker bedst for tidsserier med historisk data af tidligere sæsoner.

Prophet performer bedst på tværs af use cases.

Kilder

https://medium.com/@richa.mishr01/anomaly-detection-in-seasonal-time-series-where-anomalies-coincide-with-seasonal-peaks-9859a6a6b8ba
https://towardsdatascience.com/anomaly-detection-time-series-4c661f6f165f
https://neptune.ai/blog/anomaly-detection-in-time-series

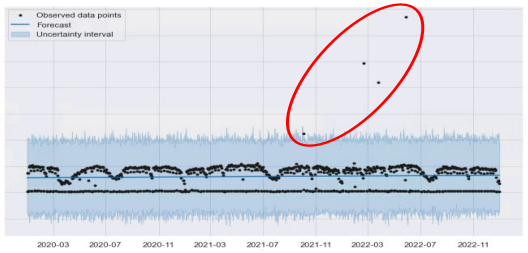
Én løsning på tværs af 3 forskellige områder



Én løsning for identificering af unormale udsving

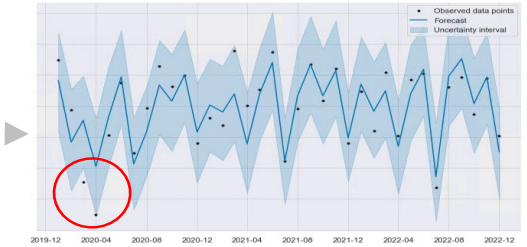
1 Ekstreme outliers

Identificer og fjern ekstreme outliers



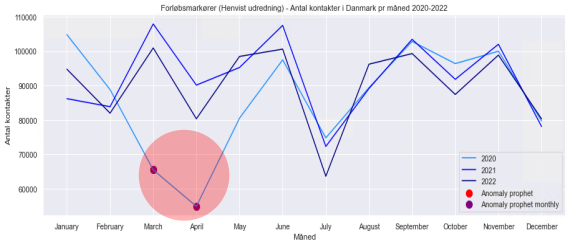
3 Udsving – månedlig data

Identificer unormale udsving baseret på månedlig data

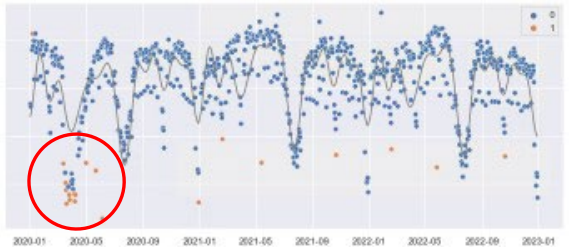


5 Færdigt resultat

Gruppér information til en visualisering

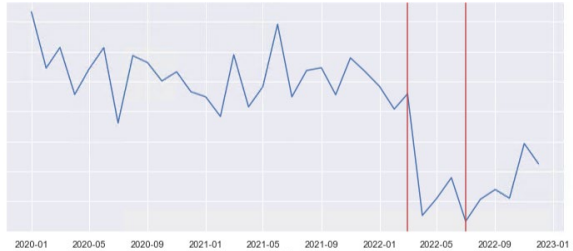



```
Antal ekstreme outliers fjernet: 4

Antal unormale udsving baseret på daglig data: 1

Antal unormale udsving baseret på månedlig data: 1

Ingen unormale changepoints identificeret
```



2 Udsving - daglig data

Identificer unormale udsving baseret på daglig data



4 Trend changepoints

Identificer trend changepoints

SUNDHEDSDATA-STYRELSEN

Værdi til områdeansvarlig

Bedre kendskab til data

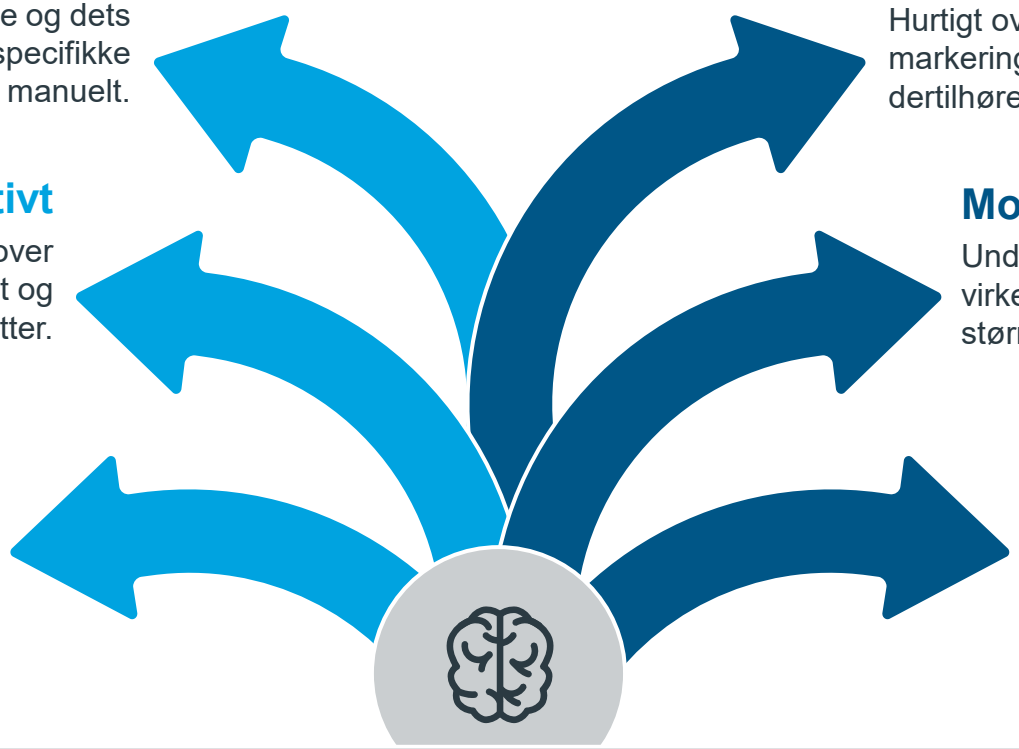
Bedre overblik over dataområde og dets datakvalitet. Finder andre og mere specifikke udsving som ikke findes manuelt.

Handle proaktivt

Giver områdeansvarlig tidligt overblik over udsving og mulighed for at handle proaktivt og tidligt kontakte indberetter.

Opfylde mindstekrav og fremtidige EHDS krav

Identificerer alvorlige mangler (fx nedbrud) og brugbar for fremtidige EHDS krav.



Færre manuelle timer

Hurtigt overblik til områdeansvarlig med markeringer af udsving som skal evalueres samt dertilhørende visualiseringer.

Monitorere større datamængder

Undgår afgrænsede regler som udelukkende virker på få datakilder. Mulighed for at monitorere større datamængder med samme proces.

Ensrettet proces

Løsning som kan facilitere en ensrettet proces for datakvalitetsmonitorering på tværs af indberetningsområder.

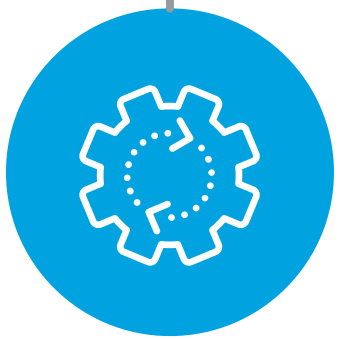
Datakvalitetsløsning – identificering af unormale udsving med AI

Perspektiver og videre anvendelse

Foreløbige resultater og perspektiver for videre anvendelse

AI har et betydeligt potentiale i Sundhedsdatastyrelsen og vil kunne anvendes på tværs af organisationen i analysearbejdet med de centrale sundhedsregistre

AI metoder kan identificere væsentlige fejl og mangler i indberetninger - også for indberetningsområder hvor der er høj datakvalitet

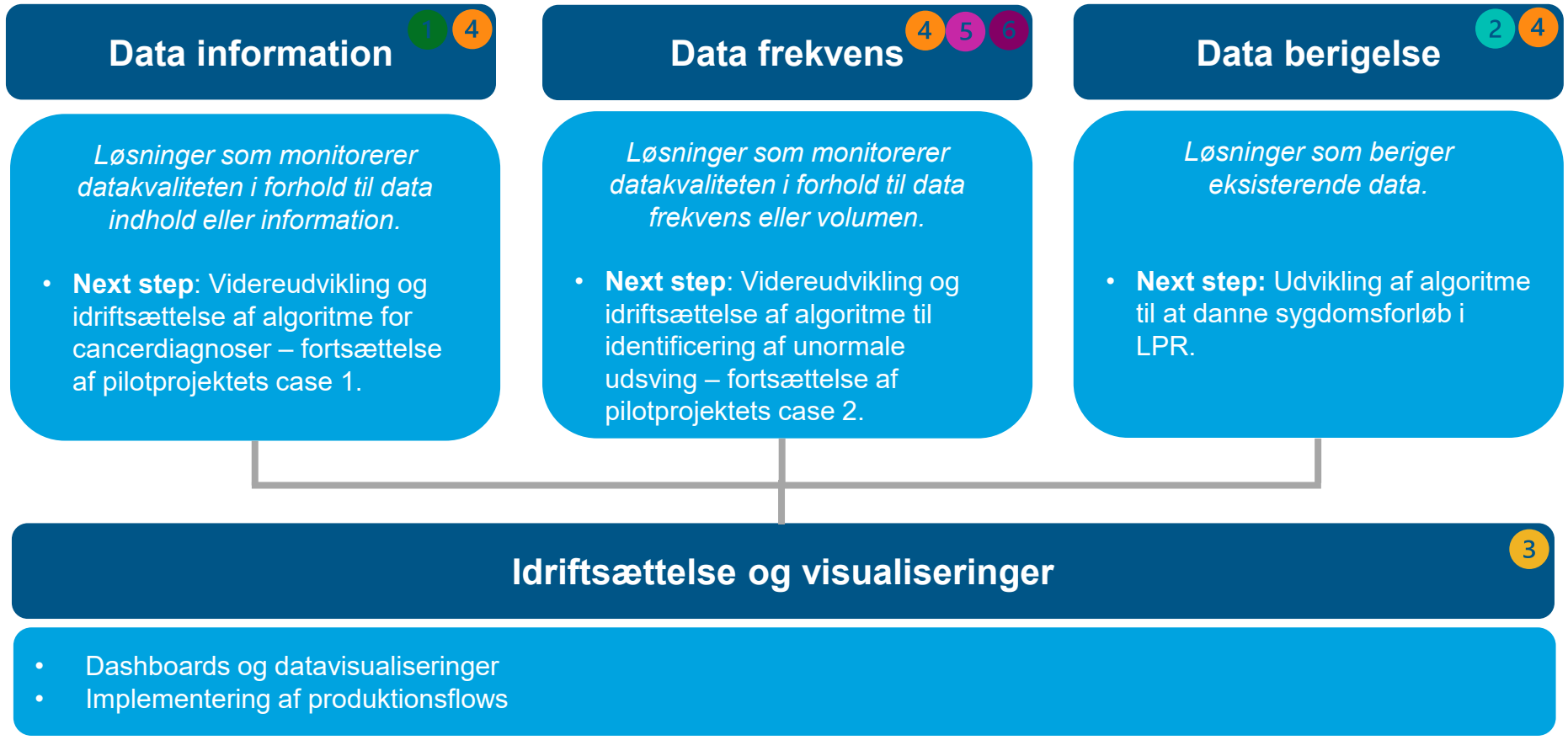


AI kan automatisere tidskrævende og komplicerede datakvalitetsopgaver

Det er muligt at opbygge kompetencer for at anvende AI baserede metoder til dataarbejdet i styrelsen

Det kræver målrettet arbejde og ressourcer at opbygge AI kompetencer og infrastruktur

Idékatalog – med EHDS datakvalitetsdimensioner



Spørgsmål