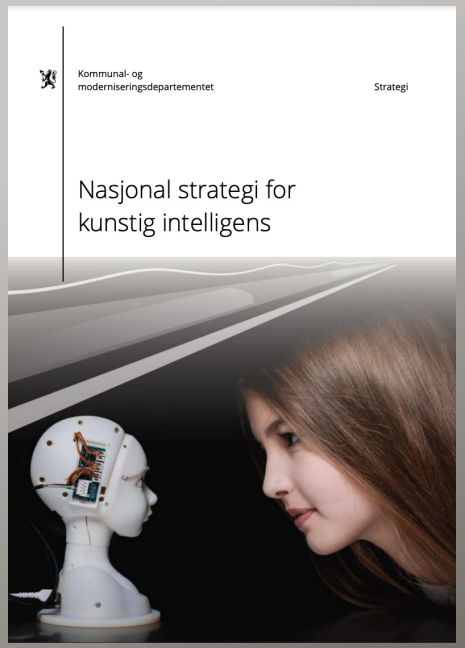
Erfaringer fra Datatilsynets regulatoriske sandkasse for kunstig intelligens

24.03.2022 // Cathrine Pihl Lyngstad (Dataseksjonen)

KI representerer et stort potensiale som vi er nødt å utforske!

Treffsikre beslutninger

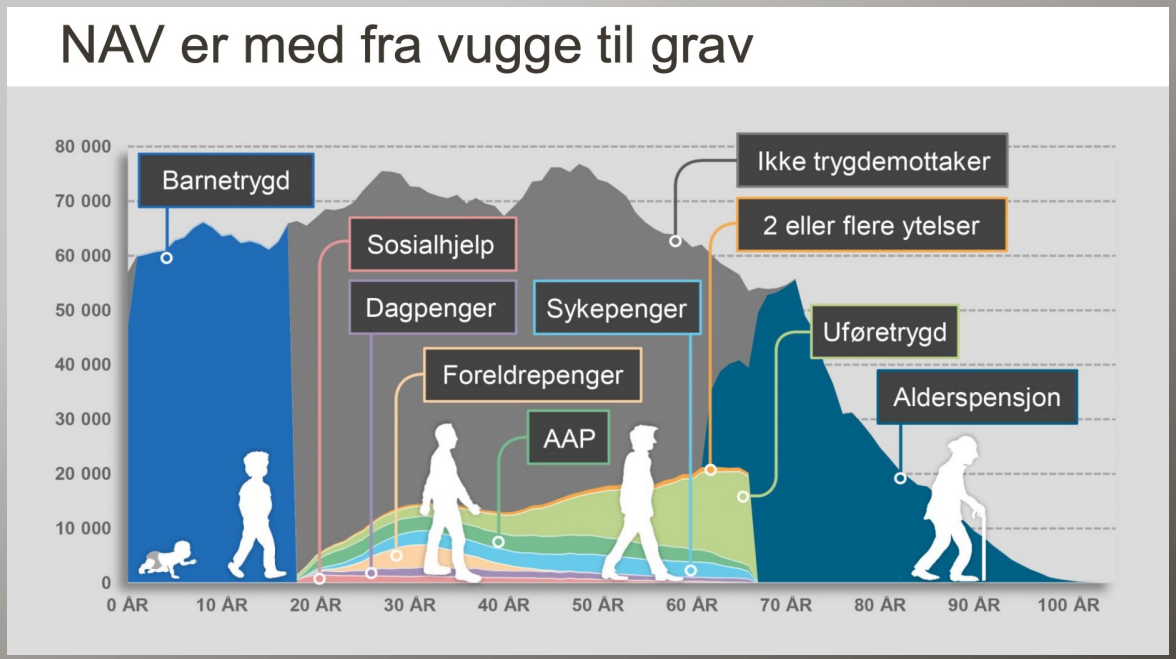


Effektivisere

Redusere risiko

Samfunnsnytte

...men det må gjøres på en ansvarlig måte...



Datatilsynet etablerte den regulatoriske sandkassen for KI og personvern i 2021, for å bidra til et godt fundament for etisk og ansvarlig kunstig intelligens

Den regulatoriske sandkassen

et kontrollert prosjektmiljø for virksomheter som vil eksperimentere med nye produkter, teknologier og tjenester som benytter kunstig intelligens i forbindelse med persondata under oppfølging av myndighetene

Hva er en god nok forklaring?

Hva betyr rettferdighets-prinsippet når *målet* med KI er å behandle ulikt?

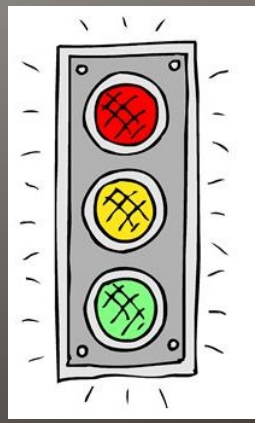
Er det begrensninger knyttet til å ta i bruk algoritmen? Ha er forenlige formål?

...

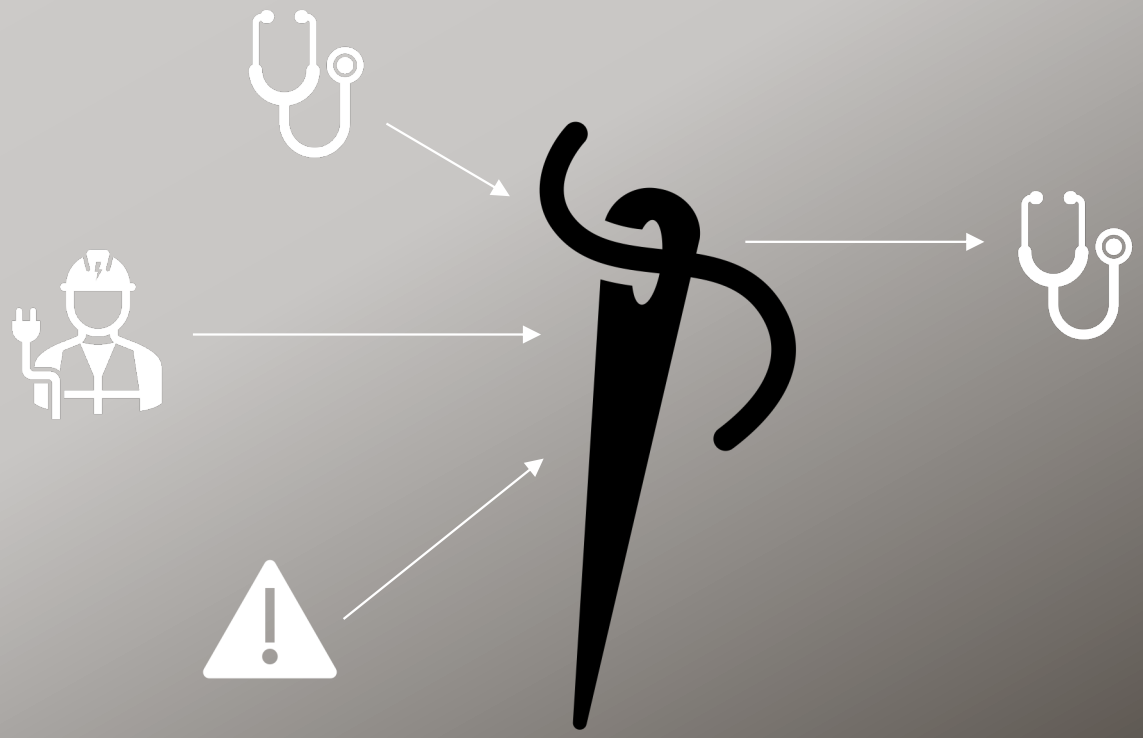
https://www.datatilsynet.no/globalassets/presentasjon-av-datatilsynets-sandkasse.pdf

Hvorfor ville NAV delta i sandkassen?

- **Kortere vei til verdi fra KI hos NAV**, gjennom raskere avklaring av juridisk handlingsrom og nødvendige tiltak/forholdsregler
- **Økt trygghet** (internt og eksternt) om vår tolkning og praksis på området
- **Styrket tillit** til NAV som ansvarlig KI-aktør
- Bruke vår posisjon til å **fremme ansvarlig bruk av KI** i Norge i stort
- Eventuell ulovlig/uansvarlig bruk av personopplysninger blir avdekket *tidlig*, og i et *trygt miljø* (ikke i formelt tilsyn)



«Sykefraværsprediksjon» kom gjennom nåløyet



SYFO-caset

Team isyfo skal bygge nytt fagsystem for sykefraværsoppfølging.De skal samtidig lage bedre tjenester for veilederne.

En av våre hypoteser er at det avholdes for mange unødvendige dialogmøter.

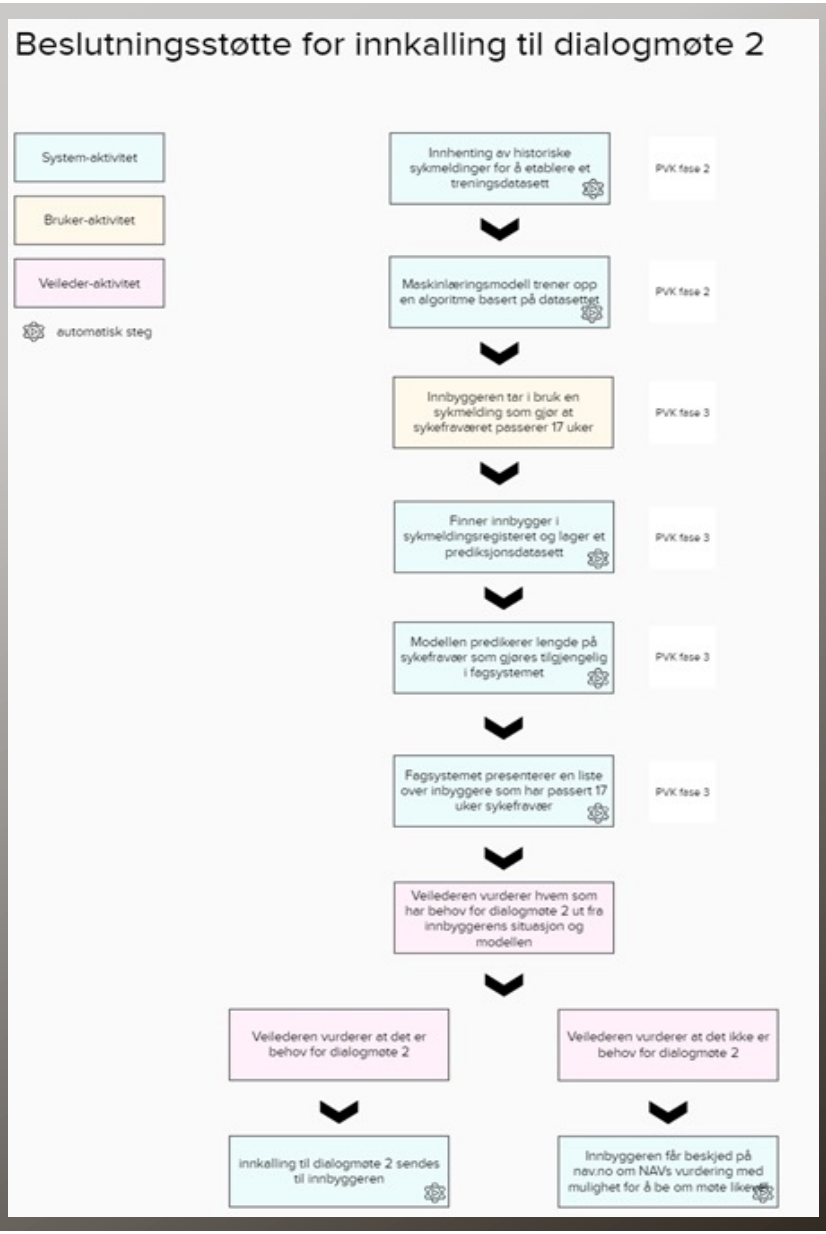
Ved å bruke maskinlæring for å predikere sykefraværslengde utover 17 uker ønsket de å understøtte veileders beslutning om behov for dialogmøte 2.

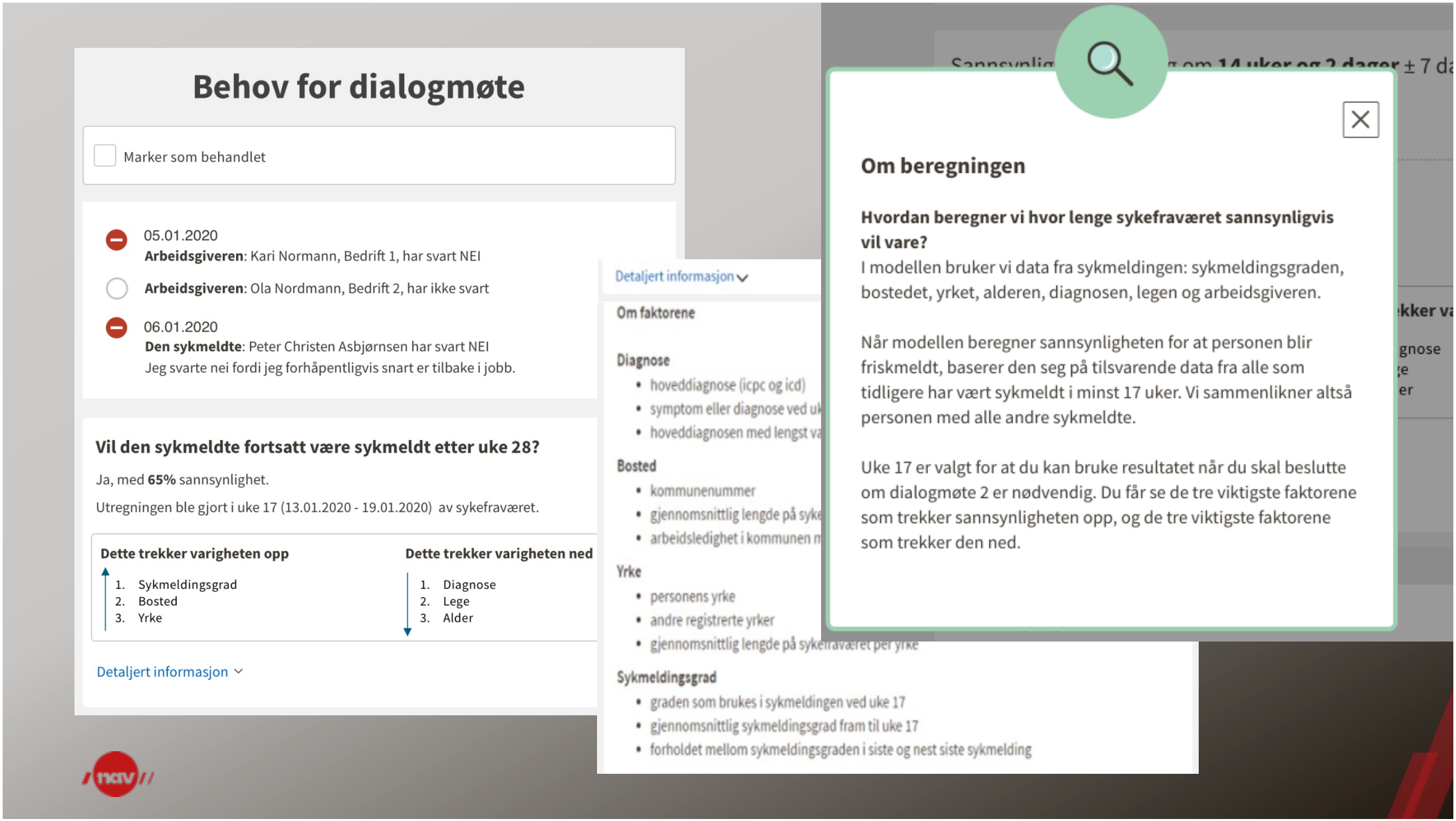
Dette ville de oppnå:

- Mer tilpasset oppfølging for sykmeldte med og uten behov for dialogmøte

- Færre unødvendige dialogmøter som sparer tid for alle involvert i sykefraværet

SYFO-caset


Beslutningsstøtte for innkalling til dialogmøte 2
System-aktivitet
Bruker-aktivitet
Veileder-aktivitet
automatisk steg
Innhenting av historiske sykmeldinger for å etablere et treningsdatasett
PVK fase 2
Maskinlæringsmodell trener opp en algoritme basert på datasettet
PVK fase 2
Innbyggeren tar i bruk en sykmelding som gjør at sykefraværet passerer 17 uker
PVK fase 3
Finner innbygger i sykmeldingsregisteret og lager et prediksjonsdatasett
PVK fase 3
Modellen predikerer lengde på sykefravær som gjøres tilgjengelig i fagsystemet
PVK fase 3
Fagsystemet presenterer en liste over innbyggere som har passert 17 uker sykefravær
PVK fase 3
Veilederen vurderer hvem som har behov for dialogmøte 2 ut fra innbyggerens situasjon og modellen
Veilederen vurderer at det er behov for dialogmøte 2
Veilederen vurderer at det ikke er behov for dialogmøte 2
innkalling til dialogmøte 2 sendes til innbyggeren
Innbyggeren får beskjed på nav.no om NAVs vurdering med mulighet for å be om møte likevel


Behov for dialogmøte
Marker som behandlet
05.01.2020
Arbeidsgiveren: Kari Normann, Bedrift 1, har svart NEI
Arbeidsgiveren: Ola Nordmann, Bedrift 2, har ikke svart
06.01.2020
Den sykmeldte: Peter Christen Asbjørnsen har svart NEI
Jeg svarte nei fordi jeg forhåpentligvis snart er tilbake i jobb.
Vil den sykmeldte fortsatt være sykmeldt etter uke 28?
Ja, med 65% sannsynlighet.
Utregningen ble gjort i uke 17 (13.01.2020 - 19.01.2020) av sykefraværet.
Dette trekker varigheten opp
1. Sykmeldingsgrad
2. Bosted
3. Yrke
Dette trekker varigheten ned
1. Diagnose
2. Lege
3. Alder
Detaljert informasjon
Detaljert informasjon
Om faktorene
Diagnose
hoveddiagnose (icpc og icd)
symptom eller diagnose ved uk
hoveddiagnosen med lengst va
Bosted
kommunenummer
gjennomsnittlig lengde på syke
arbeidsledighet i kommunen m
Yrke
personens yrke
andre registrerte yrker
gjennomsnittlig lengde på sykefraværet per yrke
Sykmeldingsgrad
graden som brukes i sykmeldingen ved uke 17
gjennomsnittlig sykmeldingsgrad fram til uke 17
forholdet mellom sykmeldingsgraden i siste og nest siste sykmelding
Om beregningen
Hvordan beregner vi hvor lenge sykefraværet sannsynligvis vil vare?
I modellen bruker vi data fra sykmeldingen: sykmeldingsgraden, bostedet, yrket, alderen, diagnosen, legen og arbeidsgiveren.
Når modellen beregner sannsynligheten for at personen blir friskmeldt, baserer den seg på tilsvarende data fra alle som tidligere har vært sykmeldt i minst 17 uker. Vi sammenlikner altså personen med alle andre sykmeldte.
Uke 17 er valgt for at du kan bruke resultatet når du skal beslutte om dialogmøte 2 er nødvendig. Du får se de tre viktigste faktorene som trekker sannsynligheten opp, og de tre viktigste faktorene som trekker den ned.

Faseinndeling for data science/KI

Rettferdighet

Modellen *skal* forskjellsbehandle, men alle modeller vil fra tid til annen gjøre feil:

- Når feilraten er større i en del av befolkningen enn i en annen, vil dette resultere i systematisk skjeve utfall for denne gruppen. I verste fall diskriminering.

- Feil gjøres også i dag, men ML kodifiserer feilraten og setter den i system. **Vi må derfor på forhånd bestemme hvilke feil vi kan leve med og hvilke vi må unngå**.

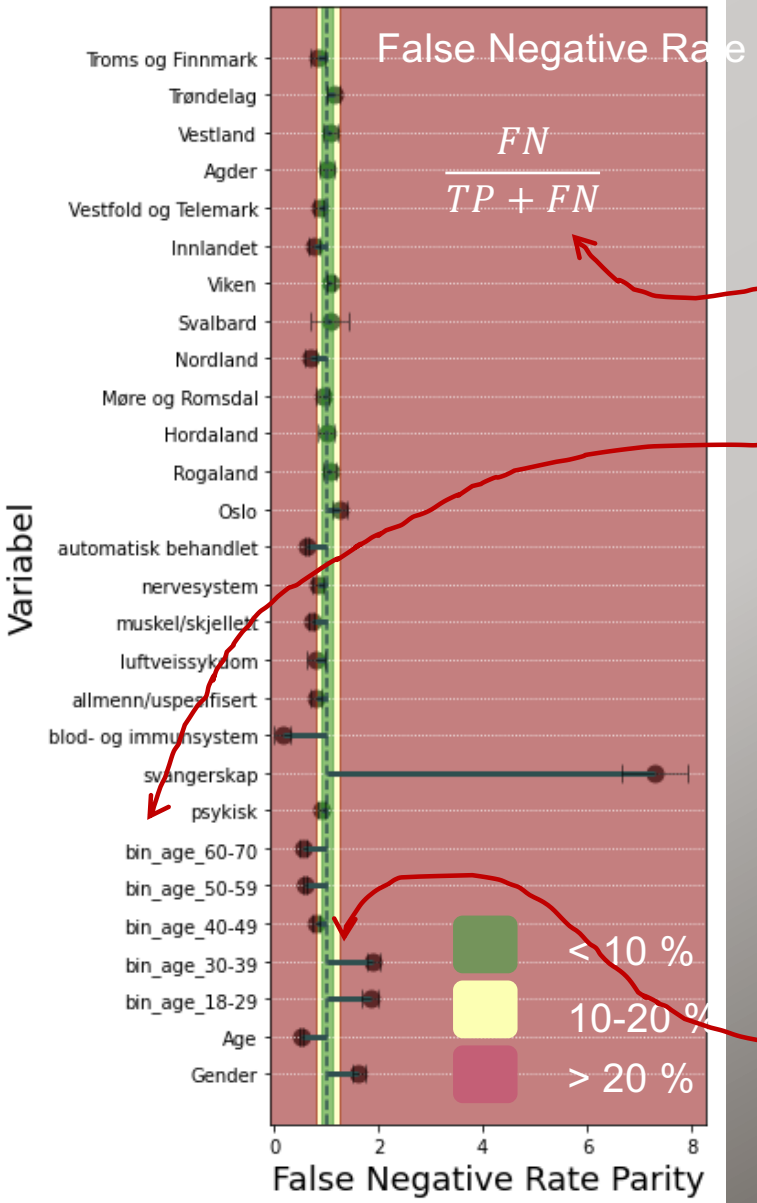
Rettferdighet er et personvernprinsipp:

*Behandlingen [] skal gjøres i **respekt** for de registrertes **interesser** og **rimelige forventninger**. Behandlingen skal dessuten være **gjennomsiktig** og **forståelig** for de registrerte, den skal ikke foregå på **fordekte** eller **manipulerende** måter.*

[Datatilsynet]

Men hvordan skal dette prinsippet forstås i lys av ML-modeller?

EDPS:
- Registrertes forventninger
- Behandlingens bredere etiske problemstillinger
- Respekt for rettigheter og friheter
- *Ikke-diskriminering*



1

Hvordan skal vi praktisk gå frem for å vurdere om modellen er rettferdig?

- Vi la frem et forslag til vurdering: en teknisk operasjonalisering/kodifisering av juridiske og etiske prinsipper med utgangspunkt i *Folketrygdloven §8-7a*

2

Hvem skal vi prioritere vurderinger for? Hvem har krav på (særlig) beskyttelse mot skjeve utfall?

- Vesentlig overlapp mellom sårbare grupper i PVF og LDL
- LDO: sammensatte diskrimineringsgrunnlag

3

Kan vi benytte markører som ikke inngår i modellen for å vurdere?

4

Hvor mye skjevhet kan vi tåle før det er å regne som diskriminering?

Rettferdighet - konklusjon

2 **Rettferdighet.** Det er viktig forskjell mellom å benytte opplysninger som allerede inngår i modellen, og å ta i bruk nye opplysninger som ikke brukes i modellen, til å sjekke for diskriminerende utfall. Det oppstår en spenning mellom personvern og rettferdighet når metoden for å avdekke og motvirke diskriminering fordrer mer *behandling av personopplysninger*.

Forklarbarhet

Hvordan forklare både modellen og enkeltprediksjoner på en måte som harmonerer med lovkrav?

Åpenhet er et personvernprinsipp

Forklarbarhet kan forstås som en operasjonalisering av dette prinsippet.

Tradisjonelt:

- Tilstrekkelig å vise hvordan PO blir brukt

...med KI:

- Kaste lys over modellens *indre logikk*
- Betydning og forventede konsekvenser for den registrerte

Mange ulike interessentgrupper har behov som relateres til forklarbarhet/forklaringer...

Data scientist

Forstå modellens oppførsel, for å kunne:

a) forbedre modellens treffsikkerhet
b) skape tillit til modellens robusthet og stabilitet

Beslutningstaker («man in the loop»)

Tillit til at prediksjonene/ anbefalingene fra algoritmene generelt er gode.

Forstå hvilke **faktorer som i størst grad påvirker prediksjonen**, og hvilken **usikkerhet** som er knyttet til, for å kunne benytte det som beslutningsstøtte.

Ledelse (ansvarlig)

Grunnleggende forståelse for «logikken» i modellen, for å

a) stole på at modellen vil fungere som forventet
b) Forstå risikoer og begrensninger

Borger

Ønsker (og har ofte krav på) en **forklaring på viktig beslutning** som er fattet om vedkommende.

Ønsker å forstå **generelt hvordan systemet virker**, for å ha tillit.

Tilsynsmyndighet

Trenger **bred og dyp dokumentasjon** for å **forstå i hvilken grad ulike krav er etterlevd,** herunder rettferdighet, hvilke data som påvirker beslutninger mm.

Trenger kode, modellmetrikker, beslutningslogger mm.

Matematiske metoder utviklet i et forsknings-samarbeid med Norsk Regnesentral

BigInsight

1

Hvor godt harmonerer vår løsning med lovkrav?

- Forordningens fortale (pkt 58):
  «*man bør etterstrebe at informasjonen som gis er meningsfull, fremfor å bruke kompliserte forklaringsmodeller basert på avansert matematikk og statistikk*»
- Visualisere tidligere utfall

Profilering =

«enhver form for ***automatisert behandling av personopplysninger*** som innebærer å bruke personopplysninger for å **vurdere visse personlige aspekter** knyttet til en fysisk person, **særlig** for å analysere eller **forutsi aspekter** som gjelder nevnte fysiske persons arbeidsprestasjoner, økonomiske situasjon, **helse**, personlige preferanser, interesser, pålitelighet, atferd, plassering eller bevegelser»

(PVF art. 4)

veileder

1

Hvor godt harmonerer vår løsning med lovvkrav?

3 **Åpenhet**. For at modellen skal gi ønsket verdi, er det avgjørende at NAV-veilederne stoler på algoritmen. Innsikt og forståelse i modellens virkemåte er viktig for å vurdere prediksjonen på et selvstendig og trygt grunnlag, uavhengig av om den endelige avgjørelsen blir å følge prediksjonens anbefaling eller ikke.

bruker

2

Art. 22: Automatisering eller profilering?

- Modellen gjør en *de facto* profilering.
- Den registrerte har derfor rett til å protestere mot at det i det hele tatt gjøres en prediksjon av sykefraværslengde.
- Plikter å gi nok informasjon til at brukeren kan ivareta sine rettigheter. Vi må forklare modellen til bruker før vi predikerer

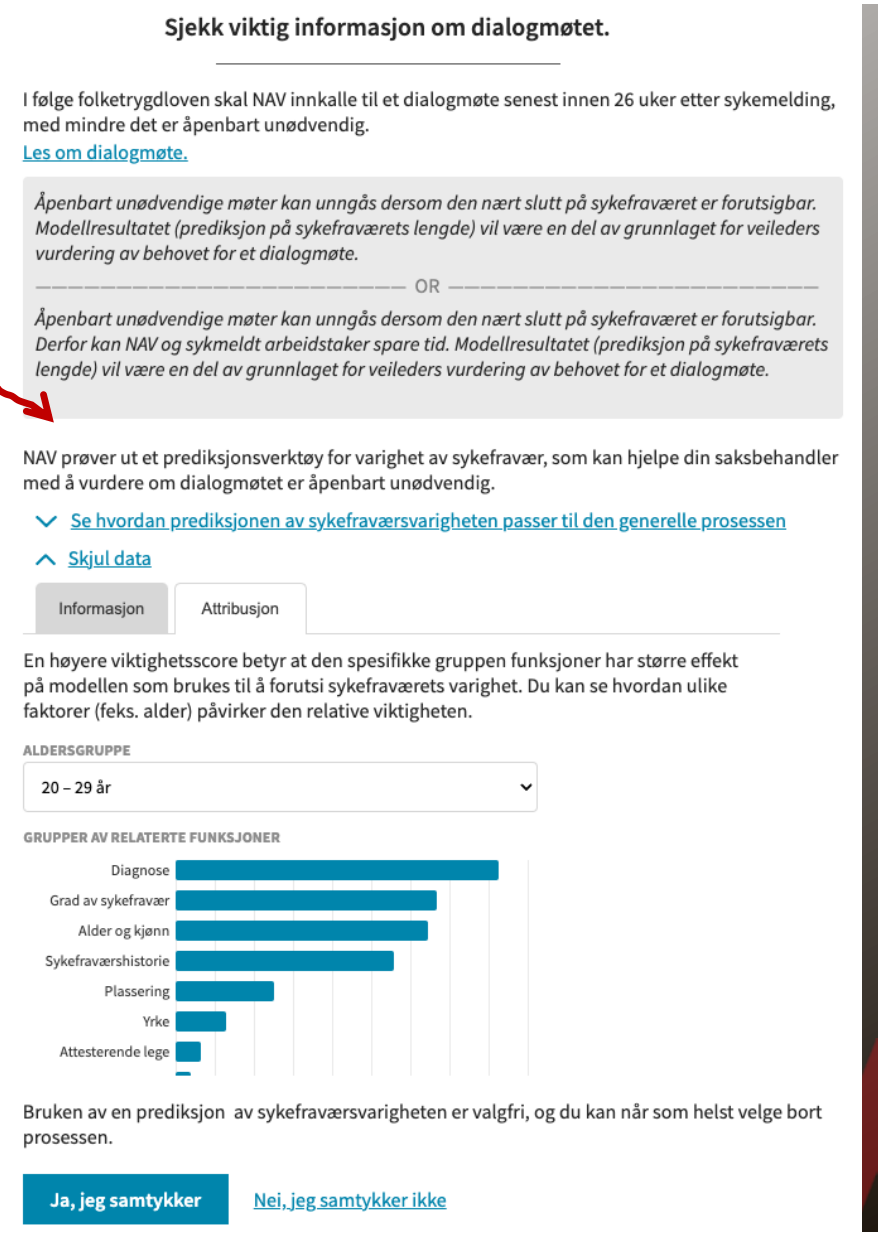
bruker

2

Art. 22: Automatisering eller profilering?

I samarbeid med UiA, UiO og NTNU er vi i gang med å prototype en retningsgivende ***template*** for modellforklaringer til bruker med bl.a. innspillene fra sandkassa.

(f.eks i tilfeller hvor art. 22 kommer til anvendelse)



"NAVs systematiske arbeid med å utvikle en modell som imøtekommer kravene til rettferdighet og forklarbarhet, viser at offentlige virksomheter godt kan være pådrivere for en ansvarlig utvikling på KI-feltet."

(Datatilsynet, Regulatorisk sandkasse, sluttrapport)

Juridisk vurdering av casen i sandkassen

Beslutnings-støtte

- Ftrl. § 8-7a er en skjønnsmessig bestemmelse
- Avgrensning mot helautomatisk behandling og Nav-loven § 4a.

Fase-inndeling

- Den juridiske vurderingen ble delt inn i ulike faser.
- Innsikt, utvikling av algoritmen, **pilot/testfase på reelle sykefraværssaker** (kontrollert bruk) og til slutt produksjonssetting.

Helhetlig vurdering

- Vurderingen av behandlingsgrunnlag gjøres innledningsvis, men lovligheten av behandlingen må sees i sammenheng med rettferdighet og forklarbarhet.

Lovlighet - konklusjon

- Utvikling
- Bruk

Konklusjoner

1 **Lovlighet.** NAV har rettslig grunnlag for å bruke KI som støtte ved beslutning om enkeltindividers behov for oppfølging og dialogmøte. Det er usikkert om det rettslige grunnlaget åpner for å bruke personopplysninger til å utvikle selve algoritmen.

Lovlighet – anvendelsesfasen

Vår inngang:

- Er det behov for eksplisitt rettslig grunnlag for selve metoden?
  - Nei, NAV kan bruke algoritme som grunnlag for beslutningstøtte jf. § ftrl § 8-7a. Sett i sammenheng med ftrl. § 21-4 og fvl. § 17.
- KI vil kunne bidra til større grad av likebehandling, bedre avgjørelsesgrunnlag og bedre ressursbruk.
  - Behandlingen er nødvendig og forholdsmessig

Hva fikk vi ut i henhold til caset og mer generelt?

- **Anbefaling om at det utarbeides særskilt lovhjemmel for utviklingsfasen. Supplerende rettslig grunnlag finnes for anvendelsesfasen, men for å ta steget videre til denne fasen mener tilsynet at det bør foreligge et mer eksplisitt hjemmelsgrunnlag for å utvikle modellen**
- Krevende juridisk øvelse, lite rettskilder og sammenlignbare caser.
- Finnes tungtveiende argumenter for og mot, hensynet til den registrertes mulighet til å ivareta sine **interesser**, samt **rettferdighet-** og **åpenhetsprisnippet**, skal imidlertid tillegges særlig vekt.
- **Likebehandling, bedre avgjørelsesgrunnlag og bedre ressursbruk er relevante juridiske argumenter ved vurdering av behandlingsgrunnlag i offentlig forvaltning.**

Betydning for andre tilfeller

Konkret vurdering i det enkelte tilfelle

- **Vanskelig å trekke mange paralleller til andre caser som ikke er veldig/helt like**

Vurderingstema

På den ene siden:

- **Særlig kategorier**
- **Mengden opplysninger («bredde» og «lengde»)**
- **Opplysninger om andre enn den som er syk**

På den andre siden :

- **Mer enhetlig behandling**
- **Mer treffsikker vurdering enn når den er manuell**
- **Mindre arbeidsbelastning i en allerede travel hverdag**

Arbeid med hjemmel

- **I dag har vi ikke et tydelig nok hjemmelsgrunnlag. Med en hjemmel er målet å klargjøre rammene som ligger der i dag + litt til.**
- **Ny hjemmel vil fortsatt ikke gi en «blankofullmakt», men gjøre det enklere å komme i mål med vurderingene.**
- **Tverrfaglig sammensatt gruppe jobber med forslag til ny hjemmel.**

Prinsipper for ansvarlig KI i NAV

 # verdi og samfunnsnytte

Vi søker nytte for brukere og samfunnet, og vi vurderer ulike konsekvenser.

 # respekt

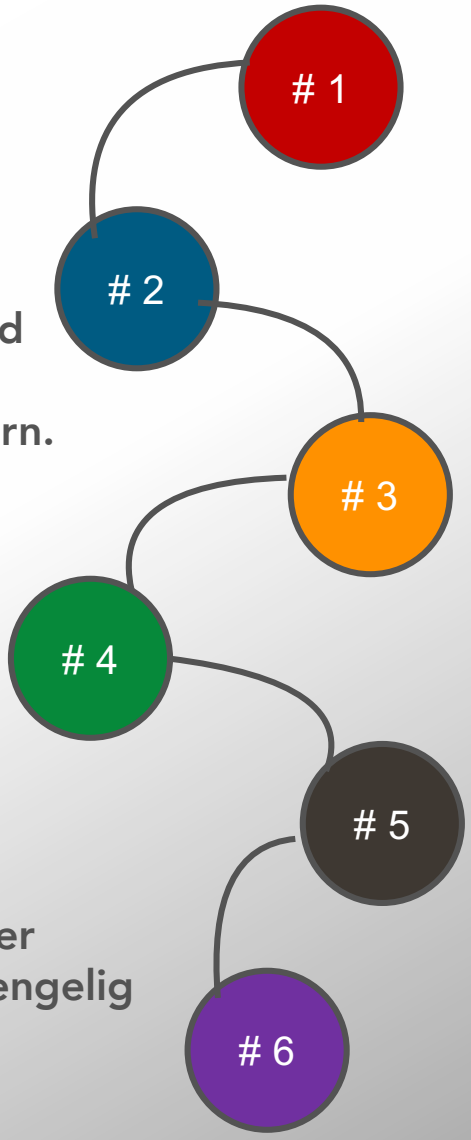
Vi behandler våre brukere med verdighet, og vi respekterer deres data og deres personvern.

 # fairness

Vi søker rettferdighet og likebehandling.

 # ansvarlighet

Vårt arbeid, både suksess og feil, er reproduserbart og kan stilles tilgjengelig for ekstern revisjon.



 # transparens

Vi er åpne og transparente, og gir minst like gode forklaringer som i dagens løsninger.

 # sikkerhet

Vi har en risikobasert og tverrfunksjonell tilnærming til sikkerhet - både modellsikkerhet og datasikkerhet.

Takk for meg!

cathrine.pihl.lyngstad@nav.no