NAV

Sluttrapport fra sandkasseprosjektet med NAV

Temaer: rettslig grunnlag, rettferdighet og forklarbarhet

Januar 2022

Datatilsynet

Innhold

1. Sammendrag

Mål med sandkasseprosjektet

NAV ønsker å bruke maskinlæring til å forutse hvilke sykmeldte brukere som vil ha behov for oppfølging to måneder frem i tid. Dette skal hjelpe veilederne med gjøre mer treffsikre vurderinger, som igjen skal spare NAV, arbeidsgivere og de sykmeldte for unødvendige møter. Målet med dette sandkasseprosjektet var å avklare lovligheten ved bruk av kunstig intelligens (KI) i denne sammenhengen, og utforske hvordan profileringen av sykmeldte kan gjøres på en rettferdig og åpen måte.

Konklusjoner

- **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.
- **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.
- **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.

Veien videre

Arbeidet med NAVs prediksjonsmodell for sykefravær, har synliggjort en stor og viktig utfordring for offentlige virksomheter som ønsker å ta i bruk kunstig intelligens: Lovene som hjemler behandling av personopplysninger, er sjeldent utformet på en måte som åpner for at personopplysningene kan brukes til maskinlæring i utvikling av kunstig intelligens. Det er viktig at lovgiver i tiden fremover legger til rette for utvikling av KI i offentlig sektor innenfor ansvarlige rammer.

Dersom NAV skal utvikle modellen videre, vil det være nødvendig med et klart og tydelig supplerende rettsgrunnlag gjennom lovgivning. En lovprosess, med tilhørende høringsrunde og utredninger, vil kunne bidra til å sikre demokratisk forankring for utvikling og bruk av kunstig intelligens i offentlig forvaltning. 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.

2. Om prosjektet

NAV har en hypotese om at det holdes for mange unødvendige møter, som stjeler tid fra arbeidsgivere, sykmeldere (f. eks leger), sykmeldte og NAVs egne veiledere. Det var motivasjonen for å etablere KI-prosjektet som skulle ta for seg prediksjon av sykefraværsvarighet.

Disse møtene er ett av flere lovpålagte stoppunkter i NAVs sykefraværsoppfølging. Innen syv ukers sykefravær skal den sykmeldte og arbeidsgiveren gjennomføre et dialogmøte. Ved åtte uker er NAV pålagt å kontrollere om den sykmeldte er i aktivitet, eller om den kan unntas fra aktivitetsplikten. Og innen et sykefravær passerer 26 uker, er NAV pålagt å vurdere behovet for et nytt dialogmøte med den sykmeldte, arbeidsgiveren og sykmelder. NAV må allerede i uke 17 lande på om et nytt dialogmøte vil være nødvendig, altså om den sykmeldte vil bli friskmeldt innen uke 26 eller ikke. På hvert av disse stoppunktene vurderer NAV hvilken type oppfølging den sykmeldte har behov for.

Dette prosjektet tar utgangspunkt i stoppunktet ved 17 ukers sykefravær, og beslutningen om å innkalle til dialogmøte 2. Ved å bruke maskinlæring for å predikere sykefraværslengden, ønsker NAV å understøtte veileders beslutning om nødvendigheten av å kalle inn til dialogmøte 2. Håpet er å:

- Redusere tidsbruk på vurdering av behov for dialogmøte for de som jobber med sykefravær hos NAV.
- Spare tid for partene som er involvert i sykefraværet, ved å i større grad unngå å kalle inn til unødvendige dialogmøter.
- Gi bedre oppfølging for de sykmeldte som har behov for dialogmøte, ved å konsentrere innsatsen mot dem som virkelig trenger det.

Presentasjon: Her er et tidlig eksempel fra NAV på hvordan en anbefaling fra systemet skal kunne presenteres for veilederen, som så skal ta den endelige avgjørelsen om det bør kalles inn til dialogmøte eller ikke. Svaret blir begrunnet med tre faktorer som taler for lengre varighet og tre som taler for kortere varighet. NAV-veilederen får også informasjon om hvordan arbeidsgivere og sykmeldte vurderer behovet for dialogmøte.

3. Sandkassemål

NAV kom inn i sandkassa med et KI-verktøy så å si klart til bruk, og med grundige juridiske vurderinger i bagasjen. Det lå til rette for at sandkasseprosjektet ville bli mer en kvalitetskontroll av utført arbeid enn en felles innovasjonsprosess. Det overordnede målet for sandkasseprosjektet var å bidra til å bygge praksis for hvordan NAV sikrer kontroll og ansvarlighet gjennom et KI-utviklingsløp. Prosjektet vil:

- Klargjøre NAVs muligheter for å benytte KI der dette er lovlig og ansvarlig.
- Korte ned vei fra idé til implementert KI på andre områder i NAV, samt for andre virksomheter som ser potensialet i lignende KI-anvendelser.

Prosjektet kan med andre ord ha nytte og overføringsverdi for NAV generelt, men også for andre virksomheter, særlig i offentlig sektor.

I sandkassa har vi diskutert problemstillinger knyttet til rettslig grunnlag, altså om NAV har lov til å bruke maskinlæring slik de planlegger. Videre har vi drøftet modellens rettferdighet, inkludert hvordan diskriminering kan avdekkes og motvirkes i en slik modell. Til sist har vi sett på hvilke krav som stilles til en meningsfull forklaring av modellen, både på system- og individnivå.

4. Vurderinger og konklusjoner

4.1 Problemstillinger

Arbeidet i sandkassen har kretset rundt tre problemstillinger knyttet til KI: rettslig grunnlag, rettferdighet og forklarbarhet. I første del ser vi på de juridiske utfordringene knyttet til NAVs rettslige grunnlag, altså lovligheten av å behandle personopplysninger for å utvikle og bruke en maskinlæringsmodell. Andre del er en vurdering av NAVs tilnærming til kravet om at en slik modell skal kunne sies å være rettferdig, og i siste del diskuterer vi problemstillinger knyttet til åpenhet og hvordan modellens virkemåte og utfall skal forklares.

4.2 Rettslig grunnlag

4.2.1 Innledning

Offentlige myndigheter behandler mange personopplysninger, og behandlingen er ofte hjemlet i lov eller forskrift. Det vil si at myndigheten slipper å hente inn samtykke eller lage avtale med hver og en den behandler personopplysninger om, men får tillatelse – et rettslig grunnlag – til å gjøre det gjennom særlover og forskrifter.

Ny teknologi kan medføre nye måter å behandle personopplysninger på, som ikke ble tatt høyde for da lovene som regulerer NAVs behandling av personopplysninger ble utformet. Utvikling og bruk av kunstig intelligens krever behandling av store mengder data – ofte personopplysninger – som sammenstilles og analyseres i en skala som ikke er mulig med andre hjelpemidler.

Det kreves tydelige lovhjemler for utvikling av kunstig intelligens i det offentlige Disse hensynene blir forsøkt ivaretatt gjennom kravene til klare hjemler i personvernforordningens artikler 5, 6 og 9, Grunnloven § 102 og Den europeiske menneskerettighetskonvensjonen artikkel 8, i tillegg til rettspraksis knyttet til disse bestemmelsene.

4.2.2 Generelt om rettslig grunnlag

Det rettslige grunnlaget som er mest aktuelt å vurdere for NAVs prediksjonsmodell, er artikkel 6-1 e. Den sier at personopplysninger kan behandles dersom det er nødvendig for å utøve offentlig myndighet, som den behandlingsansvarlige er pålagt. I tillegg kreves grunnlag etter artikkel 9 om man behandler særlige kategorier personopplysninger. Det gjør NAVs prediksjonsmodell, og dette gjelder spesielt helseopplysninger. NAV bruker derfor artikkel 9-2 b, som gir grunnlag for behandling av særlige kategorier personopplysninger for utøvelse av trygderettslige plikter og rettigheter.

Både artikkel 6-3 og artikkel 9-2 b krever et supplerende rettsgrunnlag i nasjonal rett. Det trenger ikke være en eksplisitt eller spesifikk hjemmel for den nøyaktige behandlingen. F*ormålet* med behandlingen må følge av nasjonal rett *eller* være nødvendig for å utøve offentlig myndighet.[1]

Lovhjemmelen må likevel være klar nok til å sikre forutsigbarhet for de berørte, og hindre vilkårlighet i offentlig myndighetsutøvelse.[2] Dette krever at loven definerer hvordan opplysningene kan bli brukt, og setter grenser for hvordan myndighetene kan bruke opplysningene. Det må vurderes konkret om bestemmelsen er tilstrekkelig for den aktuelle behandlingen. Jo mer inngripende behandlingen er, jo tydeligere bør hjemmelen være.

4.2.3 NAVs supplerende rettsgrunnlag

NAV bygger på supplerende rettsgrunnlag i folketrygdloven § 8-7 a, sett i sammenheng med § 21-4 i samme lov og forvaltningsloven § 17. I tillegg har NAV en hjemmel til behandling av personopplysninger i lov om arbeids- og velferdsforvaltningen (NAV-loven) § 4 a første ledd.

Folketrygdloven § 8-7 a regulerer noen av NAVs plikter til å følge opp sykmeldte. I § 8-7 a andre ledd er det regler om dialogmøte 2 som skal holdes i uke 26 av sykefraværet – unntatt «når et slikt møte antas å være åpenbart unødvendig».

Bestemmelsen må sees i sammenheng med den generelle bestemmelsen i lovens § 21-4. Den gir NAV en generell hjemmel til å samle inn opplysninger for å utøve sine oppgaver. Som forvaltningsorgan omfattes NAV også av den generelle bestemmelsen i forvaltningsloven § 17. Den krever at «forvaltningsorganet skal påse at saken er så godt opplyst som mulig før vedtak treffes».

Utviklingsfasen

Det er naturlig å dele spørsmålet om rettslig grunnlag i to, basert på de to hovedfasene i et KI-prosjekt; utviklingsfasen og anvendelsesfasen. De to fasene benytter personopplysninger på ulike måter.

I utviklingsfasen bruker NAV en stor mengde historiske data – personopplysninger om tidligere sykmeldte – fra mange registrerte, for å trene opp en modell som skal predikere andre, fremtidige personers sykefraværslengde. I utviklingsfasen benyttes det ikke personopplysninger fra de som framtidig skal få oppfølging

Spørsmålet blir dermed om de aktuelle bestemmelsene i loven (folketrygdlovens § 8-7a og § 21-4), som gir hjemmel til å behandle personopplysninger for å vurdere om det er åpenbart unødvendig å kalle inn til dialogmøte 2 i en konkret sak, også åpner for behandling av personopplysninger til utvikling av et KI-verktøy til bruk i saksbehandlingen?

En naturlig forståelse av ordlyden tilsier at disse bestemmelsene ikke gir slik hjemmel. Sammenlignet med dagens vurderinger av dialogmøte 2, vil utvikling av prediksjonsmodellen behandle et langt større volum av personopplysninger som tilhører personer som ikke lenger er sykmeldte. Et viktig moment er også at disse opplysningene i stor grad vil være særlige kategorier personopplysninger, slik som diagnose, sykefraværshistorikk og informasjon fra fritekstfelt i sykmeldingen.

Den inngripende karakteren til behandlingen i utviklingsfasen taler også for at det må kreves en klar og tydelig hjemmel. Det er tvilsomt om folketrygdloven § 8-7 a, jf. § 21-4 og forvaltningsloven § 17 er spesifikke nok til å utgjøre et tydelig og klart supplerende rettsgrunnlag etter art. 6-1 e og artikkel 9-2 b. Det kommer ikke tilstrekkelig frem i lovene NAV bygger

---

[1] Personvernforordningen artikkel 6-3

[2] jf. Grunnloven § 102 og EMK artikkel 8

på som supplerende rettsgrunnlag at opplysningene til tidligere brukere skal kunne brukes til utvikling av kunstig intelligens.

Anvendelsesfasen

For anvendelsesfasen har NAV gjort en grundig vurdering av supplerende rettsgrunnlag for bruk av prediksjonsmodellen som beslutningsstøtte. Vurderingen bygger på supplerende rettsgrunnlag i folketrygdloven § 8-7 a, sett i sammenheng med § 21-4 og forvaltningsloven § 17. I tillegg har NAV en hjemmel til behandling av personopplysninger i NAV-loven § 4 a første ledd.

NAV har vurdert at det ikke kreves særlig hjemmel i lov for selve fremgangsmåten, herunder bruk av prediksjonsmodell, men at det må gjøres en vurdering av om fremgangsmåten er forholdsmessig, for å kunne angi om den sykmeldte skal kalles inn til dialogmøte 2 eller ikke.

Avgjørende for denne vurderingen er om bruken av prediksjonsmodellen kan anses som mer inngripende for brukeren. Videre er det også foretatt en vurdering av om den planlagte bruken av personopplysninger, både når det gjelder volum og hvordan opplysningene brukes, kan anses som nødvendig for å oppfylle kravet som loven stiller.

NAV har lagt til grunn at behandlingen av personopplysninger er både forholdsmessig og nødvendig for å oppnå formålet, og vil derfor kunne ha supplerende rettsgrunnlag for å bruke prediksjonsmodellen som beslutningsstøtte i selve anvendelsesfasen, forutsatt at de har et rettslig grunnlag for utviklingen.

4.2.4 Konklusjon om rettslig grunnlag

Etter vår vurdering kan NAV ha rettslig grunnlag for å behandle personopplysninger ved anvendelse av KI i denne sammenhengen. Det er imidlertid tvilsomt om det rettslige grunnlaget som NAV har oppgitt, kan utgjøre et rettslig grunnlag for å bruke personopplysninger til å utvikle en prediksjonsmodell, selv om modellen senere skal bidra til bedre oppfølging av sykmeldte. Rettslig grunnlag for utviklingen og den tilhørende behandlingen av personopplysninger er en forutsetning for at NAV skal kunne bruke prediksjonsmodellen som beslutningsstøtte i avgjørelser som gjelder om dialogmøte 2 skal holdes.

Det kan argumenteres for at det er samfunnsmessige fordeler med at NAV kan utvikle kunstig intelligens for å forbedre og effektivisere sitt arbeid. Samtidig er utvikling av kunstig intelligens en prosess som utfordrer flere viktige personvernprinsipper. For å sikre de registrertes rettigheter, vil tydelige og klare lov- eller forskriftshjemler for en slik utvikling være nødvendig. En lovprosess, med tilhørende høringsrunde og utredninger, vil bidra til å sikre en demokratisk forankring for utvikling og bruk av kunstig intelligens i offentlig forvaltning.

Konklusjonen over er basert på diskusjonene Datatilsynet og NAV har hatt i sandkasseprosjektet, og er derfor veiledende og ikke en avgjørelse fra Datatilsynets side. Ansvaret for å vurdere det rettslige grunnlaget for de aktuelle behandlingene, ligger hos NAV som behandlingsansvarlig.

4.2.5 Automatiserte beslutningsprosesser

Selv om en behandling er lovlig, gir personvernforordningen den registrerte rett til å ikke være gjenstand for automatiserte, individuelle avgjørelser, altså avgjørelser tatt uten menneskelig inngripen, dersom behandlingen har rettsvirkning for eller på tilsvarende måte i betydelig grad påvirker den enkelte.[3] Den menneskelige involveringen må være reell, og ikke være fingert eller illusorisk.[4]

Dersom prediksjonsmodellen kun brukes som en beslutningsstøtte, vil prediksjonen om sykefraværslengden inngå som ett av flere momenter i NAV-veiledernes vurdering av om den sykmeldte skal innkalles til dialogmøte. Den menneskelige vurderingen fører i slike tilfeller til at behandlingen ikke defineres som helautomatisert. Det kan likevel tenkes at beslutningen i praksis blir helautomatisert. Veiledernes arbeidsbelastning og kunnskap om algoritmen, samt prediksjonenes opplevde og faktiske treffsikkerhet, vil påvirke risikoen for at mennesket i loopen – veilederen – på autopilot aksepterer ethvert resultat fra prediksjonsmodellen.

---

[3] Personvernforordningen artikkel 22-1

[4] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" s. 20-21.

Flere tiltak kan redusere denne risikoen. Gode rutiner og opplæring av veilederne vil være sentralt. Informasjonen som de får i forbindelse med bruk av verktøyet, må være forståelig og sette dem i stand til å vurdere prediksjonen opp mot andre momenter. Det må også innføres rutiner for å avdekke om avgjørelsene blir helautomatisert.

NAV ønsker riktignok på sikt å helautomatisere prosessen rundt innkalling til dialogmøte 2. Det finnes unntak fra forbudet mot helautomatiserte avgjørelser, men det forutsetter at avgjørelsen ikke har «rettsvirkning for eller på tilsvarende måte i betydelig grad påvirker vedkommende».[5] Så har modellen det?

NAVs prediksjonsmodell, som skal anslå lengden på sykefravær, innebærer profilering[6] og er en automatisert behandling. Så lenge modellen reelt blir brukt som beslutningsstøtte, er ikke selve avgjørelsen automatisert. Det er avgjørelsen om å kalle inn til dialogmøte eller ikke, som har potensial til å ha rettsvirkning eller tilsvarende påvirkning på den registrerte, ikke prediksjonen isolert.

Spørsmålet blir da om innkallingen til dialogmøte 2 har rettsvirkning, eller på tilsvarende måte påvirker brukeren. En avgjørelse har «rettsvirkning» dersom den påvirker personens juridiske rettigheter, slik som retten til å stemme eller kontraktsrettslige virkninger. En innkalling til dialogmøte omfattes ikke av dette. Da gjenstår det å vurdere om avgjørelsen som gjelder dialogmøte påvirker brukeren på en betydelig måte, tilsvarende som en rettsvirkning.

Svaret blir ja dersom avgjørelsen har potensial til å påvirke den enkeltes omstendigheter, adferd eller valg, har en langvarig eller permanent påvirkning, eller fører til ekskludering eller diskriminering.[7] Avgjørelser som påvirker noens økonomiske omstendigheter, slik som tilgang til helsetjenester, vil kunne kalles en påvirkning tilsvarende en rettsvirkning.

Avgjørelse om dialogmøte 2 er ikke et enkeltvedtak, men det vil kunne argumenteres for at det i «betydelig grad påvirker», og i en helautomatisert utgave vil kunne falle inn under art. 22. I tilknytning til offentlig virksomhet vil det ikke bare være enkeltvedtak som faller inn under art. 22, noe som har støtte i forarbeidene til ny forvaltningslov. Hva som faller inn under «rettsvirkning» eller «betydelig grad påvirker», må vurderes konkret ut fra hvilke konsekvenser avgjørelsen har for den registrerte. For NAVs prediksjonsmodell kan det tenkes et skille mellom situasjoner der det innkalles til dialogmøte 2, og der det ikke innkalles til møte.

Dersom den sykmeldte ikke blir kalt inn til dialogmøte, oppstår det ingen plikt for vedkommende. Samtidig beholder den sykmeldte retten til å kreve et dialogmøte. I slike situasjoner vil avgjørelsen ha mindre inngripende virkning på den registrerte, så lenge muligheten til å be om dialogmøte er reell. Samtidig skal dialogmøte 2 hjelpe den sykmeldte i å komme tilbake i jobb. Ikke alle sykmeldte vil ha ressurser til å benytte seg av retten til å be om dialogmøte. Dette kan kanskje delvis ivaretas med god informasjon til de registrerte.

I de situasjonene en sykmeldt blir innkalt til dialogmøte 2 – som er hovedregelen etter folketrygdloven § 8-7 a – vil det oppstå en plikt for den sykmeldte til å møte opp. Manglende etterlevelse av denne plikten, vil i ytterste konsekvens kunne føre til bortfall av sykepenger. I slike tilfeller vil plikten til å møte på dialogmøte 2 kunne ha stor påvirkning på den sykmeldte, og vil kunne falle inn under vilkåret i artikkel 22.

Kort oppsummert kan avgjørelser om å kalle inn til dialogmøte kunne nå opp til terskelen i artikkel 22, noe som utløser et forbud. Avgjørelser om ikke å kalle inn vil kunne falle utenfor terskelen, forutsatt at den sykmeldtes rett til å be om dialogmøte er reell. Om det er praktisk mulig å skille avgjørelsene på denne måten, vil være opp til NAV å vurdere.

---

[5] Personvernforordningen artikkel 22-1

[6] Personvernforordningen artikkel 4-4

[7] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" s. 21-22

4.3 Rettferdighet

4.3.1 Innledning

Når vi i dette sandkasseprosjektet har diskutert rettferdighet, har vi tatt utgangspunkt i tre hovedprinsipper for ansvarlig kunstig intelligens: lovlig, etisk og sikker. Disse er hentet fra «Retningslinjer for pålitelig kunstig intelligens» som er utarbeidet av en ekspertgruppe oppnevnt av EU-kommisjonen. De samme prinsippene er også gjenspeilet i Nasjonal strategi for kunstig intelligens.

EDPB lister i sin veiledning om innebygd personvern opp flere momenter som inngår i rettferdighetsprinsippet, blant annet ikke-diskriminering, den registrertes forventninger, behandlingens bredere etiske problemstillinger og respekt for rettigheter og friheter.[8] Rettferdighetsprinsippet inneholder altså flere elementer enn ikke-diskriminering. Diskriminering i algoritmer er en kjent utfordring ved kunstig intelligens, og sandkassearbeidet har derfor sentrert seg rundt dette. En stor offentlig aktør som NAV, har et særlig ansvar for å være bevisst den skjeve maktbalansen i møte med brukere av systemene deres.

Rettferdighetsprinsippet står sentralt i andre lovverk, blant annet ulike menneskerettighetsbestemmelser og likestillings- og diskrimineringsloven. Disse lovverkene vil også kunne få betydning for spørsmålet om rettferdighet, og kan også tenkes å gå lengre eller kortere i kravene sine enn personvernreglene.

4.3.2 NAVs modell

NAV har utviklet metoder som gjør at de kan teste hvor rettferdig modellen er. Hovedfokuset har vært modellens *bias,* altså potensielle skjevheter i datainnsamling, valg av variabler, modellvalg eller implementering, og hvordan disse kommer til uttrykk i skjeve utfall og eventuelle diskrimineringseffekter. Maskinlæringsmodeller vil uunngåelig behandle mennesker ulikt ettersom ønsket om en mer brukertilpasset forskjellsbehandling ofte motiverer utviklingen av en maskinlæringsmodell. Hvordan unngå usaklig forskjellsbehandling, var ett av de sentrale temaene i dette sandkasseprosjektet. NAV ønsker ikke å reprodusere eller befeste eksisterende skjevheter, men risikerer å gjøre nettopp det hvis skjevhetene ikke blir analysert og adressert.

For å understøtte en slik analyse, ønsker NAV å gjøre en vurdering av hva et rettferdig algoritmeutfall innebærer i rettslig forstand. Å utvikle en maskinlæringsmodell som imøtekommer flere lovverks[9] krav til rettferdighet, innebærer en operasjonalisering av juridiske og etiske prinsipper.

For å vurdere om modellen er forenelig med rettferdighetsbegrepene i lovverket, er det nyttig å sannsynliggjøre hvordan modellen vil oppføre seg når den er satt i produksjon. Hvilke utfall kan for eksempel grupper med særlig krav om beskyttelse mot urettmessig diskriminering forvente å få?

NAV peker selv på at en slik analyse ikke er dekkende for alle måter behandlingen av personopplysninger kan være urettferdig eller diskriminerende. Men ved å fokusere på utfallet (uavhengig av forhold knyttet til for eksempel datainnsamling, -prosessering og praktisk modellanvendelse), muliggjør den en diskusjon av hvordan rettferdighetsbegrepet skal forstås og hvordan det kan operasjonaliseres.

I operasjonaliseringen av rettferdighetsvurderingen har NAV valgt å fokusere på utfallsrettferdighet, altså hvorvidt utfallet av modellen fordeler seg rettferdig på tvers av ulike grupper. Vurderingen er komparativ, altså ser den på hvordan ulike grupper som inngår i modellen behandles sammenlignet med hverandre, og ikke målt opp mot en standard eller norm. NAV har også vurdert modellfeil som kaller inn til dialogmøte der det ikke er nødvendig, som mindre alvorlig enn det motsatte. Ett av utgangspunktene for å vurdere rettferdighet i prediksjonsmodellen, er folketrygdens § 8-7a, som instruerer NAV til å holde et dialogmøte «unntatt når et slikt møte antas å være åpenbart unødvendig». Et slikt krav antyder at det i tvilstilfeller heller bør holdes ett dialogmøte for mye enn for lite.

Fra et personvernperspektiv må rettferdighet vurderes både på gruppenivå og individnivå. Modellen vil kunne være i strid med rettferdighetsprinsippet også dersom kun individer blir påvirket negativt i betydelig grad, og ikke bare dersom

---

[8] Guidelines 4/2019 on Article 25 Data Protection by Design and by Default | European Data Protection Board (europa.eu)

[9] I tillegg til personvernforordningen må NAV forholde seg til forvaltningsloven, NAV-loven og likestillings- og diskrimineringsloven

det skjer en gruppevis diskriminering – for eksempel dersom det er sjeldne kombinasjoner av faktorer som fører til svært negative virkninger for den registrerte.

I tillegg kan det tenkes at prediksjon av sykefraværslengde for enkelte grupper vil slå feil ut når det gjelder vurdering av om det skal innkalles til dialogmøte. Dette kan for eksempel gjelde i tilfeller der framtidig lengde på sykefravær ikke er det beste vurderingsmomentet for avgjørelse av om dialogmøte er «åpenbart unødvendig», og hvor man ut fra et rettferdighetsperspektiv muligens må identifisere slike typetilfeller for å unngå en slik ubalanse. Det kan for eksempel tenkes at flere gravide har lange sykefravær der det fortsatt er åpenbart unødvendig med dialogmøte 2. Det samme kan muligens gjelde for delvis uføretrygdede som skal sykmeldes i ett år fra sin resterende arbeidsprosent med et framtidig mål om full uføretrygd.

4.3.3 Andre momenter

Modellen som er blitt diskutert i sandkassa er et beslutningsstøttesystem. Det betyr at prediksjonen vil være ett av flere informasjonselementer som går inn i veilederens vurdering. Ved en eventuell helautomatisert beslutning, bør det gjøres en ny rettferdighetsvurdering. Samtidig er det viktig å huske på at også mennesker diskriminerer. Det er derfor ikke gitt at det faktiske utfallet for den registrerte blir mer rettferdig av at det er et menneske i loopen. Likevel kan det oppleves som mer inngripende å bli urettferdig behandlet av en maskinlæringsmodell enn av en veileder. I tillegg vil modellens eventuelle urettferdige praksis skalere på en helt annen måte enn dagens system og føre til systematisert urettferdighet. En ny vurdering av den registrertes berettigede/rimelige forventninger til behandlingen, vil sannsynligvis bli enda viktigere i en helautomatisert modell. Det gjelder også revisjon og kontroll av algoritmene.

4.3.4 Hvem har krav på særlig beskyttelse?

Metoden som er valgt for å evaluere maskinlæringsmodellens utfallsrettferdighet, krever at NAV definerer hvilke grupper som skal evalueres opp mot hverandre. I utgangspunktet finnes det vilkårlig mange brukergrupper som kan defineres ut ifra brukermassen som utgjør datagrunnlaget for trening av modellen. Hvilke brukergrupper som skal inngå i en rettferdighetsvurdering av modellen, er et spørsmål med flere ulike sosiale, historiske og samfunnsmessige dimensjoner. NAV er til for alle, men det er verken teknisk eller praktisk mulig å gjøre en vurdering for alle gruppeidentiteter i det norske samfunnet. Hvem som har krav på eller særlig behov for beskyttelse mot skjeve modellutfall, blir dermed et sentralt spørsmål.

Store deler av dette spørsmålet faller mer naturlig inn under likestillings- og diskrimineringsloven, og som en del av sandkassearbeidet inviterte vi inn likestillings- og diskrimineringsombudet for å drøfte disse spørsmålene.

I utgangspunktet er gruppene NAV opererer med – blant annet kjønn, alder og diagnoser – godt forankret i likestillings- og diskrimineringsloven. Det kan tenkes at det i tillegg til de definerte gruppene, også vil oppstå sammensatte diskrimineringsgrunnlag, hvor en kombinasjon av gruppetilhørighet slår spesielt skjevt ut. Det finnes også andre sårbare grupper som det kan være nyttig å inkludere, slik som rusavhengige, personer med omsorgsoppgaver og personer med lav økonomisk status.

Et sentralt spørsmål knyttet til diskriminering, er om en slik prediksjonsmodell forskjellsbehandler på en slik måte at det kan kalles diskriminering. Siden den konkrete modellen som vurderes omhandler sykefraværslengde, og er knyttet til hvor vidt det skal kalles inn til et dialogmøte eller ikke, når man ikke nødvendigvis denne diskrimineringsterskelen. Det vil sannsynligvis stille seg annerledes med en modell for andre typer ytelser med større konsekvenser for den registrerte.

4.3.5 Spenningen mellom personvern og rettferdighet

I alle maskinlæringsmodeller kan det oppstå spenning mellom modellens virkemåte og flere personvernprinsipper. I NAV-prosjektet oppstår et slikt spenningsforhold når NAV skal oppfylle plikten sin til å sjekke om modellen behandler skjevt eller diskriminerer. I utgangspunktet må man behandle personopplysninger både for å avdekke og for å korrigere utfallsskjevhet. Avdekking av skjevhet i modellens utfall, kan riktignok gjøres uavhengig av om gruppetilhørigheten er en del av modellen. Men for å gjennomføre en evaluering av modellens utfall, må gruppetilhørigheten brukes. Til slutt kan det være mulig å tilfredsstille andre krav til informasjonsrettferdighet uten slik behandling av personopplysninger.

Disse spørsmålene er sentrale for utviklere av ansvarlig KI, og forslaget til ny KI-lovgivning fra EU berører spørsmålene.[10]

NAVs tjenester skal være tilgjengelige for hele befolkningen, og NAV må derfor navigere spenningsforholdet mellom personvern og skjeve utfall i hver modell som utvikles. I tillegg er det en stor overlapp mellom gruppene som personvernforordningen definerer som sårbare og gruppene som omfattes av likestillings- og diskrimineringsloven.

Når modellens rettferdighet skal vurderes, er det fra et personvernståsted forskjell på det å benytte opplysninger som allerede inngår i modellen og det å ta i bruk nye opplysninger som i utgangspunktet ikke benyttes i modellen, men som legges til analysen for å sjekke for diskriminerende utfall. Det oppstår en slik spenning mellom personvern og rettferdighet, når metoden for å avdekke og motvirke diskriminering fordrer omfattende behandling av særskilte kategorier av personopplysninger. Opplysninger som allerede er inkludert i algoritmen er en del av beslutningsgrunnlaget i sykefraværsoppfølgingen. Helt nye opplysninger er derimot avhengig av en ny lovlighetsvurdering. I tillegg er det sannsynlig at de registrerte har en berettiget forventning om at opplysninger som er uvedkommende for vurderingen av om det skal innkalles til drøftingsmøte ikke skal brukes inn i modellen. Det kan tenkes at bruk av anonymiserte eller syntetiske data kan være en løsning, som kan avdekke utfallsskjevhet samtidig som personvernet ivaretas. Fullt ut anonymiserte data regnes ikke som personopplysninger, og dermed kommer ikke personvernforordningen til anvendelse. Dette har vi imidlertid ikke diskutert inngående i sandkassen.

Det finnes ikke nødvendigvis et fullgodt svar på spørsmålet om spenningen mellom personvern og rettferdighet i en maskinlæringsmodell. Like fullt er det en sentral del i diskusjonen om og arbeidet mot ansvarlig kunstig intelligens.

4.3.6 Et toleranserom for forskjellsbehandling?

Formålet med prediksjonsmodellen er å understøtte en form for forskjellsbehandling: å bistå veileder i vurderingen av hvem som bør få tilbud om dialogmøte. Det sentrale spørsmålet vil derfor ikke være hvor vidt modellen forskjellsbehandler, men snarere om den forskjellsbehandler korrekt, samt at forskjellsbehandlingen ikke er urimelig og/eller diskriminerende.

Modellen, som skal predikere sykefravær, er i praksis et automatisert bidrag til de mange tusen vurderingene som hver dag gjøres av veilederne i NAV. Det finnes metoder for å vurdere hvor rettferdige utfallene av en prediksjonsmodell blir, noe som gjør det mulig å tallfeste rettferdigheten på en måte som er umulig i dag. Følgelig kan man ved bruk av en maskinlæringsmodell avdekke diskriminerende utfall som i dag er skjult bak den daglige arbeidsflyten på Norges NAV-kontor. Det åpner for en vanskelig diskusjon om hvor mye urettferdighet man skal akseptere, og hvordan man forholder seg til et slikt tallfestet urettferdighet. Ingen vil påstå at alle NAV-klienter behandles rettferdig, men en maskinlæringsmodell vil nådeløst tallfeste en slik rate.

Det er neppe mulig å sette en prosentsats for et akseptert toleranserom for diskriminering slik likestillings- og diskrimineringsloven er innrettet. Hvilken praksis som fører til den reelt sett største diskrimineringseffekten er like fullt noe norske og europeiske likestillings- og diskrimineringsombud må ta stilling til i møte med slik teknologi.

[10] EU-kommisjonens forslag til ny forordning om kunstig intelligens artikkel 10-5. Hentet fra EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)

4.4 Hvordan forklare bruken av kunstig intelligens?

4.4.1 Åpenhet og forklarbarhet

Åpenhet er et grunnleggende prinsipp i personvernforordningen.[11] I tillegg til å være en forutsetning for å avdekke feil, forskjellsbehandling eller andre problematiske forhold, bidrar det til å skape tillitt og setter enkeltindividet i stand til å bruke sine rettigheter og ivareta sine interesser. I tilknytning til KI bruker man ofte begrepet forklarbarhet, som går direkte på de KI-spesifikke problemstillingene knyttet til åpenhet, og som kan sies å være en konkretisering av åpenhetsprinsippet. Tradisjonelt har åpenhet dreid seg om å vise hvordan ulike personopplysninger brukes, men bruk av KI krever andre metoder, som kan forklare komplekse modeller på en forståelig måte.

Forklarbarhet er et interessant tema, både fordi det kan være en utfordring å forklare komplekse systemer og fordi hvordan kravet til åpenhet skal implementeres i praksis vil variere fra løsning til løsning. I tillegg muliggjør maskinlæringsmodeller forklaringer som ser radikalt annerledes ut enn de vi er vant til, gjerne basert på avanserte matematiske og statistiske modeller. Dette åpner for en viktig avveiing mellom en mer korrekt, teknisk forklaring eller en mindre korrekt, men mer forståelig forklaring.

I denne delen av rapporten deler vi vurderinger og konklusjoner fra diskusjonene vi hadde rundt åpenhet og forklarbarhet i NAVs løsning for å predikere sykefraværslengde. Veiledere på NAV-kontor og den sykmeldte som individ er de to mest sentrale målgruppene for forklaring i dette tilfellet.

4.4.2 Krav til åpenhet

Uavhengig om du bruker kunstig intelligens eller ikke er det visse krav til åpenhet dersom du behandler personopplysninger.[12] Kort oppsummert er disse:

- De registrerte må få informasjon om hvordan opplysningene brukes, enten opplysningene hentes inn fra den registrerte selv eller fra andre[13]
- Informasjonen må være lett tilgjengelig, for eksempel på en hjemmeside, og være skrevet i et klart og forståelig språk[14]
- Den registrerte har rett til å få vite om det behandles opplysninger om henne og eventuelt innsyn i egne opplysninger[15]
- Det er et grunnleggende krav at all behandling av personopplysninger skal gjøres på en åpen måte. Det betyr at det er krav om å vurdere hvilke åpenhetstiltak som må til for at den registrerte skal kunne ivareta egne rettigheter[16]

I det første kulepunktet er det krav om å gi informasjon om hvordan opplysningene brukes. Det inkluderer blant annet kontaktinformasjon til den behandlingsansvarlige (i dette tilfellet NAV), formålet med behandlingen og hvilke kategorier personopplysninger som blir behandlet. Dette er informasjon som typisk formidles i personvernerklæringen.

Når det gjelder kunstig intelligens kan det være verdt å merke seg kravet om å forklare algoritmens underliggende logikk. Det er et spesifikt krav å gi «relevant informasjon om den underliggende logikken samt om betydningen og de forventede konsekvensene av en slik behandling for den registrerte».[17] Det er ikke nødvendigvis innlysende hvordan disse kravene skal forstås. Man bør etterstrebe at informasjonen som gis er meningsfull, fremfor å bruke kompliserte forklaringsmodeller basert på avansert matematikk og statistikk.[18] Det understrekes også i forordningens fortale at teknologisk kompleksitet gjør åpenhet ekstra viktig.[19] De forventede konsekvensene bør også eksemplifiseres, for eksempel ved hjelp av visualisering av tidligere utfall.

---

[11] Personvernforordningen artikkel 5-1 a og fortalepunkt 58

[12] Les gjerne mer i detalj om krav om åpenhet i KI-løsninger i rapporten Kunstig intelligens og personvern (2018)

[13] Personvernforordningen artikkel 13 og 14

[14] Personvernforordningen artikkel 12

[15] Personvernforordningen artikkel 15

[16] Personvernforordningen artikkel 5

[17] ICO, personvernforordningen artikkel 13 og 14

[18] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" s. 31

[19] Personvernforordningens fortalepunkt 58

Det spesifiseres at dette i hvert fall skal gjøres i tilfeller der det skjer automatiserte avgjørelser eller profilering etter artikkel 22. Om det må informeres om logikken dersom det ikke er automatiserte avgjørelser eller profilering, må vurderes fra sak til sak basert på om det er nødvendig for å sikre en rettferdig og åpen behandling.

4.4.3 Automatisk eller ikke?

Hvis en behandling kan kategoriseres som en automatisert avgjørelse eller profilering etter artikkel 22 stilles det ekstra krav til åpenhet.[20] Du har blant annet rett til å vite om du blir utsatt for automatiserte avgjørelser, herunder profilering. Det er også et krav at individet får relevant informasjon om den underliggende logikken, betydningen av og de forventede konsekvensene av en slik behandling, som nevnt over.

Men har du rett på en individuell forklaring om hvordan algoritmen kom frem til avgjørelsen? Selve lovteksten sier ikke det, men i fortalen står det at den registrerte har krav på en forklaring på hvordan modellen kom frem til resultatet, det vil si hvordan opplysningene er vektet og vurdert i de konkrete tilfellene, dersom man faller inn under artikkel 22.[21] Fortalen uttaler også at den registrerte bør «informeres om forekomsten av profilering og konsekvensene av dette.»[22] Fortalen i seg selv er ikke juridisk bindende og gir ikke alene en rett til en individuell forklaring.

Åpenhetskravet betyr nødvendigvis ikke at kildekoden må gjøres tilgjengelig, men forklaringen må gjøre den registrerte i stand til å forstå hvorfor en avgjørelse ble som den ble. Dette gjelder der avgjørelsen faller inn under artikkel 22 om automatiserte individuelle avgjørelser. Det kan også tenkes tilfeller der rettferdighets- og åpenhetsprinsippet stiller høyere krav til forklaring, for eksempel ved profilering som ikke oppfyller vilkårene i artikkel 22, men hvor gode grunner tilsier at den registrerte burde få slik informasjon.

En meningsfull forklaring er ikke bare avhengig av tekniske og juridiske krav, men også språklige og designmessige vurderinger. Det må også vurderes hvilken målgruppe forklaringen retter seg mot, noe som vil kunne innebære en forskjell for veiledere og brukere. Også selve den praktiske anvendelsen av forklaringsmodellen i veiledernes arbeidshverdag vil kunne innebære at tilliten og opplevelsen av om en får en meningsfull forklaring vil kunne variere, ved at forklaringene som gis fremstår som standardiserte og derfor gir liten veiledning over tid. Samfunnsmessige forhold som tillit til virksomheten, vedtakets betydning og tilliten til KI-systemer generelt vil også kunne påvirke opplevelsen av en meningsfull forklaring.

Et sentralt spørsmål for NAV har vært om prediksjonsmodellen for sykefraværslengde er en automatisert avgjørelse og utløser disse ekstra kravene eller ikke. I dette tilfellet er det liten tvil om at prediksjonsmodellen ikke er en helautomatisert behandling. Prediksjonen vil være ett av flere informasjonselementer en veileder skal vurdere før avgjørelsen tas.

Likevel finnes det grunner til å informere om logikken og virkemåten i modeller som ikke er helautomatiserte. Prediksjonsmodellen gjennomfører uansett profilering,[23] og en meningsfull forklaring bidrar til å bygge tillit og er et uttrykk for ansvarlighet. I tillegg vil en meningsfull forklaring sette veilederen bedre i stand til å vurdere hvor mye vekt hen skal gi anbefalingen algoritmen gir.

Uavhengig av om det er snakk om en helautomatisert beslutning eller ikke plikter databehandleren å gi nok informasjon til at brukeren har den informasjonen som er nødvendig for å kunne ivareta sine rettigheter. NAVs sentrale rolle i offentlig forvaltning gir opphav til en asymmetrisk maktrelasjon mellom bruker og etat, som også er et argument for å etterstrebe en så meningsfull forklaring som mulig, til tross for at modellen ikke er helautomatisert.[24]

---

20 Personvernforordningen artikkel 13-2 f og 14-2 g, se også «Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 og artikkel 4-4

21 Personvernforordningens fortalepunkt 71

22 Personvernforordningens fortalepunkt 60

23 Personvernforordningen artikkel 4-4

24 Kunstig intelligens og personvern | Datatilsynet

4.4.4 Stoler vi på algoritmen?

Gode forklaringer av algoritmen og dens prediksjoner øker tilliten hos de som skal bruke systemet, noe som er helt sentralt for å oppnå den ønskede verdien. De flere tusen NAV-ansatte som jobber med brukerveiledning spiller derfor en avgjørende rolle.

Systemet som predikerer sykefraværslengde er et beslutningsstøttesystem, men hva skjer hvis systemet i praktisk anvendelse blir et beslutningssystem? En veileder i NAV gjennomgår mange saker i løpet av en vanlig arbeidsdag. Hvis det virker som om algoritmen gir konsekvente gode anbefalinger kan det jo være fristende å alltid følge den. Veilederen tenker kanskje at algoritmen sitter på så mye data at den vet best, og at det skal litt til for å ikke følge anbefalingen? Hvor lett er det for en nyansatt å ikke følge anbefalingen til algoritmen?

Eller hva hvis veilederne syns at algoritmen gir merkelige anbefalinger og ikke stoler på dem? En konsekvens av det ville være at veilederne konsekvent ikke bruker det som beslutningsstøtte. Det ville også ha vært uheldig fordi hele hensikten med løsningen er å hjelpe veilederne til å ta gode valg, slik at innkallingene oftere treffer riktig. Ideelt sett vil en slik modell redusere de tilfeldige variasjonene blant veilederne og føre til mer enhetlig praksis, i tillegg til å redusere kostnader.

I sandkassen diskuterte vi risikoene for at veileder lener seg for mye eller for lite på beslutningsstøttesystemet, og hvordan legge til rette for at systemet oppleves som en reell støtte for veileder og blir brukt på en god og riktig måte. At veileder får god opplæring og instrukser i hvordan algoritmen fungerer og skal brukes, samt en meningsfull forklaring i enkelttilfeller, er viktig for å redusere risikoen for en «snikautomatisering» eller at den ikke tas med i vurderingen i det hele tatt. Når NAV-veilederne forstår modellens oppbygning, virkemåte og oppførsel, vil det være enklere å vurdere prediksjonen på et selvstendig og trygt grunnlag. I tillegg kan forklaringen bidra til å hjelpe veileder å avdekke diskriminering, uønsket forskjellsbehandling og feil. Her vil en forklaring knyttet til en enkelt avgjørelse være supplert med informasjon knyttet til utfallet for enkelte grupper det er naturlig å sammenlikne med.

Forklaring av modellen: Her er et eksempel på hvordan en generell forklaring om hvordan modellen opererer, kan se ut for veilederen som bruker systemet.

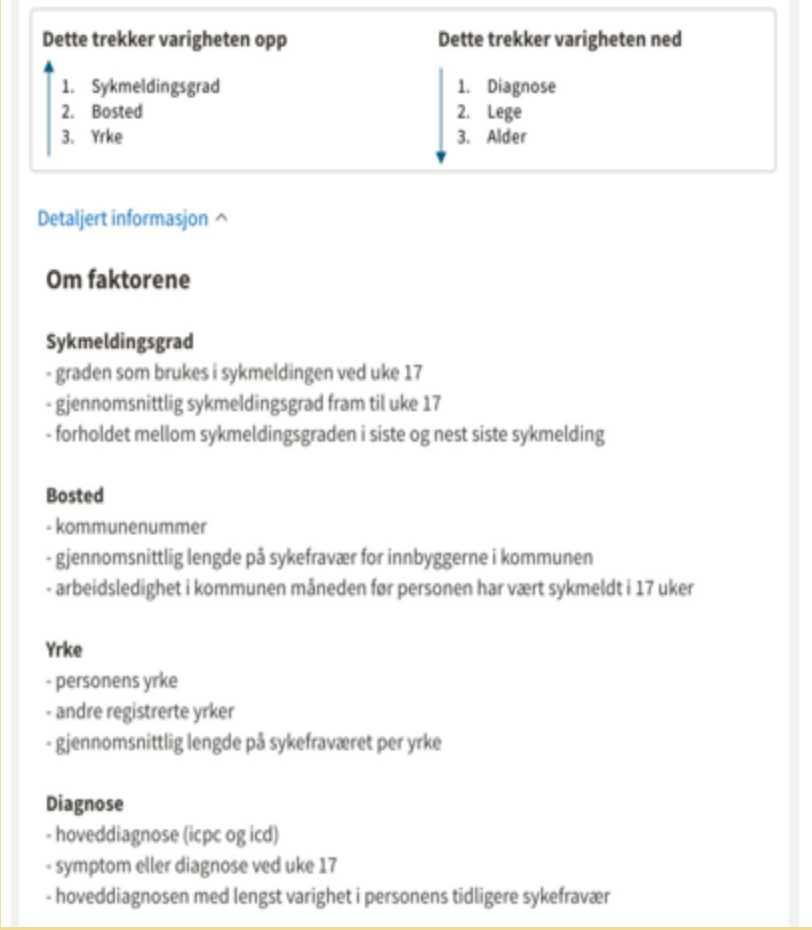
4.4.5 Hvordan ser en meningsfull forklaring ut?

Et spørsmål vi har diskutert i sandkassen er hvordan en meningsfull forklaring ser ut i praksis i NAVs tilfelle. Målgruppene for åpenhet i løsningen er de sykmeldte og NAV-veiledere. Forklaringene er både globale, altså på systemnivå, og lokale utfallsforklaringer. De to ulike nivåene vil følgelig ha delvis ulike målgrupper, og det vil stilles ulike krav til hvordan de innrettes.

NAV ønsker å informere i forkant av behandlingen om at brukeren har rett til å protestere mot at det i det hele tatt skal gjøres en prediksjon basert på en profilering. De ønsker også å informere om hvordan modellen er bygget og hvilke variabler som inngår. NAV vurderer også å informere den individuelle brukeren om de viktigste faktorene som trekker den predikerte sykefraværsvarigheten opp og de viktigste faktorene som trekker den ned.

En meningsfull forklaring er ikke bare avhengig av tekniske og juridiske krav, men også språklige og designmessige vurderinger. Forklaringen må tilpasses til målgruppen den retter seg mot. For eksempel trenger veiledere i NAV forklaringer som kan anvendes i praksis i en hektisk hverdag. NAV må derfor balansere og avveie mellom dybde og forenklinger som gjør det mulig å ta forklaringen i bruk. Forklaringen må dessuten integreres med øvrig informasjon veileder har tilgang til. Et konkret eksempel er at NAV ikke kan presentere informasjon om hvordan 100 variabler har bidratt til en prediksjon. NAV må gruppere disse sammen og gjøre et utvalg. Det kreves i tillegg ekstra årvåkenhet dersom forklaringen retter seg mot barn eller sårbare grupper. NAVs modell vil kunne inkludere flere særskilte kategorier personopplysninger om sårbare grupper og NAV vil derfor måtte vurdere å tilpasse språk, innhold og form basert på det.

Informasjon om data: Slik ser NAV for seg at man kan bli presentert hvilke data som brukes og legges til grunn på hvilke måte i modellen.



Når det gjelder veiledere planlegger NAV å forklare hvordan modellen virker generelt og beskrive hvordan de skal bruke resultatet fra modellen i saksbehandlingsrutiner. I tillegg skal veiledere få forklaringer på enkeltsaksnivå og informasjonselementer modellen har lært fra som en del av informasjonsgrunnlaget for å ta den endelige avgjørelsen om brukeren kalles inn til dialogmøtet eller ikke. Prediksjonen skal inngå som ett av flere momenter som er tilgjengelig for veilederen, inkludert den informasjonen en veileder baserer en avgjørelse på i dag.
I tillegg til de to hovedmålgruppene (brukere og veiledere) som nevnes her har NAV identifisert forretningssiden/ledelse, de ansvarlige for modellen og tilsynsmyndigheter som andre målgrupper som vil ha behov for og krav på en forklaring på hvordan algoritmen fungerer.

NAV ønsker å ta sin del av ansvaret når det kommer til åpenhet rundt bruken av algoritmer. Et mulig tiltak som diskuteres er å informere om hvordan NAV i stort ønsker å ta i bruk kunstig intelligens. NAV søker også å bidra til bred informasjon og opplyst debatt om bruken av kunstig intelligens gjennom mediebildet. Et siste tiltak er å informere og involvere brukerutvalg i forkant av og underveis i utviklingen av tjenester som baserer seg på kunstig intelligens.

5. Veien videre

Arbeidet med NAVs prediksjonsmodell for sykefravær, har synliggjort en stor og viktig utfordring for offentlige virksomheter som ønsker å ta i bruk kunstig intelligens: Lovene som hjemler behandling av personopplysninger, er sjeldent utformet på en måte som åpner for at personopplysningene kan brukes til maskinlæring i utvikling av kunstig intelligens. Det er viktig at lovgiver i tiden fremover legger til rette for utvikling av KI i offentlig sektor innenfor ansvarlige rammer.

Dersom NAV skal utvikle modellen videre, vil det være nødvendig med et klart og tydelig supplerende rettsgrunnlag gjennom lovgivning. En lovprosess, med tilhørende høringsrunde og utredninger, vil kunne bidra til å sikre demokratisk forankring for utvikling og bruk av kunstig intelligens i offentlig forvaltning. 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.

Datatilsynets regulatoriske sandkasse for ansvarlig kunstig intelligens

Besøksadresse:
Trelastgata 3, Oslo

Postadresse:
Postboks 458 Sentrum
0105 Oslo

sandkasse@datatilsynet.no
Telefon: +47 22 39 69 00

datatilsynet.no/sandkasse
personvernbloggen.no
twitter.com/datatilsynet