Analyse av COMPAS

07.05.2024

Bjarte Holgersen **Cecilie Sivertsen Rønnestad** **Sadaf Samin Dar**

Simon Lyng-Jørgensen **Åsmund Olai Sand-Larsen**

Innhold

Introduksjon

I 2016 utga ProPublica en analyse av Northpoint sin algoritme ved navn COMPAS. COMPAS brukes i det amerikanske rettsvesenet for å vurdere hvorvidt siktede skal varetektsfengsles, prøveløslates eller forvares. Algoritmen gir en risikovurdering basert på statistisk analyse av data, og dommere bruker denne vurderingen som en ressurs for å ta beslutninger i rettsvesenet.

ProPublica sin analyse argumenterer for at COMPAS diskriminerer mot fargede - altså at den gir uberettiget høye risikovurderinger til fargede kontra hvite individer. Northpoint svarer i tur på ProPublica sin analyse ved å påpeke mangler ved den, og konstatere at deres algoritme ikke er urettferdig, men at enkelte eksisterende ulikheter i verden videreføres da det er avtrykk av dem i dataen algoritmen baseres på.

I denne oppgaven skal vi gjøre en revisjon og evaluering av COMPAS algoritmen. Vi vil bruke dataene ProPublica baserte sin analyse på for å gjøre vår egen analyse. Først presenterer vi den teoretiske bakgrunnen for oppgaven ved å diskutere hvorfor algoritmer brukes, diskutere hva en rettferdig algoritme er og om man kan lage en rettferdig algoritme. Deretter vil vi presentere funn fra vår analyse av COMPAS. og hvorvidt COMPAS kan anses som en. Videre vil vi diskutere ProPublica sine funn sammenlignet med Northpointe sine, og komme med anbefalinger for tredjeparts-algoritme revisjoner i fremtiden.

Hvorfor bruke en algoritme?

Det er flere grunner til å bruke en algoritme til å vurdere risiko i rettsvesenet. Det grunnleggende formålet er at det kan hjelpe med å ta valg som er rettferdige og effektive for både rettssystemet, tiltalte og samfunnet som helhet. Bruk av algoritmer kan fremme objektivitet i risikovurderinger ved å bruke standardiserte kriterier og dataanalyse. Det er kjent at mennesker ofte kan gi forskjellige evalueringer av samme sak på grunn av endring i eksterne årsaker som blodsukker, mengde søvn eller personlig bias. Algoritmer kan også bidra drastisk til effektivitet ved å automatisere risikovurderingsprosessen. Med økt effektivitet kan man spare ressurser og ha større kapasitet i rettssystemet, noe som antagelig

kan hjelpe å fremme et rettferdig system. Rettferdighet er tross alt målet til et velfungerende rettssystem - vi ønsker at alle skal bli likt behandlet under loven.

Man kan argumentere for at algoritmer fremmer rettferdighet ved å bidra til å standardisere og objektivisere risikovurderingsprosessen. Dette er dog omstridt, ProPublica argumenterer for at den standardiserte prosessen COMPAS tilføyer, bærer med seg ulikhetene som finnes i verden. Dette får oss til å stille noen viktige spørsmål - om algoritmene våre bærer ulikhetene som finnes i verden videre, gjør den det mer eller mindre enn oss mennesker? Bør vi aktivt interferere med dette? Eller er det eventuelt en grunn til at vi ikke burde bruke disse algoritmene i det hele tatt.

Den beste grunnen til å bruke algoritmer er sannsynligvis effektiviseringen de kan føye til. Om det står mellom å ikke i det hele tatt få behandlet en sak på grunnlag av manglende ressurser, og å gjøre en ganske god algoritmisk saksbehandling, er det lettere å stille seg bak bruken av algoritmer i rettsvesenet. I en ideell verden, hvor det alltid er nok dommere til å gjøre nøye og menneskelige risikoanalyser, er det mulig at det ville vært den åpenbare veien å gå. Dette avhenger igjen om man tror at algoritmen kan oppnå høyere treffsikkerhet i prediksjonene sine, og deretter om denne treffsikkerheten er det høyeste mål for denne prosedyren.

Teori

Hvordan ser en rettferdig algoritme ut?

Hvordan vil en rettferdig algoritme se ut? Dette er et vanskelig og omfattende spørsmål å ta stilling til. Første relevante puslebrikke for å besvare et slikt spørsmål er en forståelse for hva rettferdighet betyr. Et grunnleggende og sikkert utgangspunkt er å si at rettferdighet betyr lik behandling for like tilfeller. I sammenheng med ProPublica sin sak mot COMPAS algoritmen til Northpoint betyr dette at algoritmen er rettferdig om den gir samme risiko-score til like tilfeller, uavhengig om tiltalt er farget eller hvit - noe ProPublica mente at denne algoritmen ikke gjør. Om dette stammer fra algoritmens urettferdighet eller verdens urettferdighet kan bestrides. Northpoint sier i sin respons til ProPublica at algoritmen selv er nøytral, men urettferdigheter som finnes i verden plukkes opp i dataen og videreføres.

Det er vanskelig å gi et konkret eksempel på hvordan en rettferdig algoritme skal se ut, men lettere å peke på noen idealer å strekke seg etter. Vi ønsker at algoritmen vår er forutsigbar og gjennomsiktig. Disse to henger sammen - om det er gjennomsiktig hvordan en algoritme kommer frem til resultatet sitt er det også forutsigbart hva slags output den vil gi for en vilkårlig sak. Algoritmen bør heller ikke diskriminere på grunnlag av verken rase, kjønn, religion, sosioøkonomisk status eller andre beskyttede kategorier. Dette vil være i åpenbar konflikt med rettferdighet som lik behandling av like tilfeller. I tillegg ønsker vi at algoritmen skal ha riktig balanse mellom sensitivitet (evnen til å identifisere reelle risikoer) og spesifisitet (evnen til å unngå falske positive). En rettferdig algoritme bør helst også bidra til å redusere urettferdigheter som eksisterer i verden allerede, og da ikke forsterke eller skape urettferdig ulikhet. Det er også nødvendig at algoritmen gjør gode prediksjoner om den skal brukes til å ta avgjørelser om hvordan folk skal behandles. Om en algoritme har svært dårlig treffsikkerhet er det lite å si i forsvar for å bruke den til å ta avgjørelser i et rettssystem.

I artikkelen «The algorithm audit: scoring the algorithms that score us" gir forfatterne en liste med dimensjoner for å evaluere en algoritme. De overordnede kategoriene de foreslår å evaluere er; (Brown, S., Davidovic, J., & Hasan, A., 2021)

- Partiskhet
- Effektivitet
- Gjennomsiktighet
- Direkte påvirkninger (Potensiale for misbruk, og overtredelse av rettigheter)

- Sikkerhet og tilgang

En rettferdig algoritme vil antagelig oppnå en gunstig balanse i prioriteringen mellom disse dimensjonene - da forbedring av forskjellige dimensjoner kan være gjensidig utelukkende, tvinges vi til å gjøre prioriteringer dem imellom. Her ligger den etiske dybden i spesifisere hva en rettferdig algoritme består i. Forskjellige personer vil veie disse dimensjonene forskjellig, og hvis vi i tillegg jobber med flere konsepsjoner av rettferdighet, ender vi opp med en dyp og komplisert problemstilling.

Med tanke på COMPAS algoritmen vi tar stilling til, er det et viktig spørsmål hvorvidt man vil prioritere en rettferdig mekanisme eller rettferdige resultater. COMPAS algoritmen er ubevisst etnisiteten til individene de gjør risikoanalyse for, allikevel viser analyse av dens resultater å urettferdig behandle fargede. Da kan man spørre seg om det er riktigere at vi gjør algoritmen bevisst etnisitet, og aktivt interfererer ved å regne med, og balansere ut eksisterende urettferdigheter.

Hvis det finnes eksisterende urettferdigheter i samfunnet, er det mulig for en algoritme å gi rettferdige utslag?

Flere filosofer og deltakere i samfunnsdebatten peker på ulike måter hvor algoritmer er med på å forsterke de eksisterende urettferdighetene i samfunnet. Byung-Chul Han er kritisk til teknologiens rolle i samfunnet, og det potensialet algoritmer har til å forsterke eksisterende maktstrukturer og sosiale ulikheter. Særlig kritisk er han til sosiale medier som gjennom sine algoritmer bidrar til å skape en kultur av synlighet, ytelse og konkurranse, som ifølge ham fører til sosial og psykologisk utmattelse (Han, 2015).

Et problem er at algoritmene fungerer ved å prosessere data som er hentet inn fra samfunnet, og blir på den måten formet av de iboende skjevhetene som dataen inneholder fra den konteksten de er samlet inn under. Algoritmer som er trent på data fra et samfunn preget av rasediskriminering føre med seg og forsterke disse, for eksempel gjennom ansettelsesprosesser (O'Neil, 2016).

Mye data bærer med seg historiske urettferdigheter som rasediskriminering og sosialøkonomisk ulikhet, og algoritmer brukt i system for å forutsi fremtidig kriminell adferd vil av den grunn uforholdsmessig peke ut personer med visse etniske bakgrunner eller fra områder preget av lavinntektshusholdninger (O'Neil, 2016).

Benjamin bruker begrepet «The New Jim Code», med referanse til det gamle systemet med segregering i sør-statene i USA, som er blitt kalt Jim Crow. Hun mener med dette at algoritmer og system som er laget på data produsert i et samfunn preget av etterdønningene fra et rasistisk samfunn er med på å gjenta tidligere urettferdigheter, og fortsetter med diskriminering og undertrykking (Ruha, 2019).

Hvordan kan man eventuelt gjøre opp for urettferdighet i en algoritme?

Benjamin mener at for å få algoritmer til å gi rettferdige utslag, så krever det en grunnleggende omstrukturering av hvordan data samles inn, analyseres og brukes. Dette vil blant annet innebære at utviklingen av algoritmer må være åpen for inspeksjon og for regulering. Det må være klare regler for hvordan data kan benyttes, og det må være et lovverk som faktisk er i stand til å holde selskaper ansvarlige for misbruk. Selskapene og systemene må reguleres slik at den teknologiske utviklingen ikke bare tjener kommersielle krefter men også tar sosiale hensyn og sosialt ansvar. Det krever også at de som utvikler disse systemene i langt større grad tar etiske hensyn og utformer teknologien slik at den kan brukes til å fremme sosial rettferdighet, og ikke utnytter svake grupper og forsterker sosial urettferdighet for økonomisk vinning (Ruha, 2019).

Vil det å gjøre opp for urettferdighet i en algoritme kunne føre til andre problemer?

Når det gjelder spørsmålet om det å gjøre opp for urettferdighet vil føre med seg andre problemer, er det et spørsmål som ikke nødvendigvis har med teknologi eller algoritmer å gjøre. Det handler blant annet om hvordan vil skal definere urettferdighet, og om det er mulig å gjøre opp for urettferdighet uten samtidig ramme uskyldige. Handler det om like muligheter eller om like utfall? Det kan være at algoritmer som retter opp for skjevheter i data vil føre med seg færre goder for andre grupper i samfunnet. Vil det å bruke rase som parameter for omfordeling være mer rettferdig enn sosial klasse og sosioøkonomisk status? Det er bare å peke på den polariseringen som har økt kraftig i USA de siste 10 årene, og diskusjon rundt støtte basert på rase fremfor klasse. Fattige hvite, som utgjør en stor gruppe i USA, opplever ikke at de har privilegier og goder fordi de er hvite. En økonomisk omfordeling som forfordeler hvite kan for dem oppleves meget urettferdig og slike tiltak ting kan føre til økte gnisninger mellom grupper og en økt mistillit til myndighetene (Hughes, 2024).

Jeg tenker også at det er et spørsmål om «framing».. Er rettferdighet, like rettigheter og tilgang til goder ett nullsum-spill, eller er det noe som alle kommer bedre ut av? For dem som ser på det som et nullsum-spill, så vil nødvendigvis det at en gruppe får mer enn før innebære at andre får mindre, og av den grunn vil være imot det. Men om en ikke tenker på det slik, men snarere at alle tjener på at alle blir behandlet på likefot av mennesker og teknologi, så vil prosessen dit være lettere.

Det er også et spørsmål om politisk vilje, eller flertallets vilje til å «gjøre det riktige». Jeg er av den oppfatning at «vi» som regel vet hva som er riktig å gjøre, eller hva som er rett og rimelig for flertallet å gjøre, men at det er vanskelig å gjøre det riktige. Det er bare å bruke klima- og miljøsaken som et eksempel på dette. Forskningsmiljøene sier at vi nærmer oss et «point of no return» når det kommer til klimagassutslipp, forurensing og overforbruk av jordens ressurser. Men vi evner ikke å gjøre nok med det.. ikke fordi vi ikke vet hva som bør gjøres, men fordi vi ikke vil gi slipp på de godene vi har, særlig når andre ikke gjør det samme. Vi vil ha billigere strøm, men ikke vindmøller i nærheten. Vi vil ha elbiler, men fortsatt nyte godt av oljen som pumper penger inn i kassa. Vi ønsker renere luft, samtidig som vi klager over flyprisene til ferien. La oss si at om tre år så har den kraftigste K.I-modellen laget en plan for hvordan verden skal rette klimaet, og det innebar en ny velferdsstandard som tilsvarte Norge anno 1999. Ville vi godtatt dette, eller hadde vi nektet og sådd tvil om det for å beholde våre goder?

Poenget med denne digresjonen er å peke på at det grunnleggende problemet ikke nødvendigvis er algoritmene vi bruker, men snarere samfunnet og den menneskelige natur. Det er lett å ha tiltro til et system, en algoritme, du selv drar nytte av. Altså, det kan være lett å overse problemene det fører med for andre, og vanskelig å godta en endring hvor du selv ikke blir like tilgodesett som før.

Skjevhet fra Bayes' setning

Hvis vi definerer hendelsen A som at en person gjentar forbrytelsen sin og B som at en person er afroamerikansk, får vi fra Bayes' setning denne sammenhengen:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Fra formelen ser vi at $P(A|B)$, altså sannsynligheten for at en person gjentar forbrytelsen sin gitt at den er afroamerikansk, er invers proporsjonal med $P(B)$, sannsynligheten for å være afroamerikansk. Siden afroamerikanere er den største gruppen i datasettet med omtrent 51%, vil det gjøre $P(A|B)$ liten. Da blir $P\left(\overline{A}|B\right)$, altså sannsynligheten for å ikke gjenta en forbrytelse gitt at man er afroamerikaner, høy.

Resultater

Truth table for race: African-American			
	Actual true	Actual false	Sum
Predicted true	0.3931	0.1830	0.5761
Predicted false	0.1654	0.2586	0.4239
Sum	0.5584	0.4416	