Analyse av COMPAS-algoritmen: Rettferdighetens avveininger

Fartein, Filip, Trajan og Jonatan

Universitetet i Oslo, april 2022

Abstract

I denne rapporten legger vi frem en analyse av COMPAS-algoritmen, et beslutningsverktøy som bestemmer om en tidligere innsatt kommer til å begå et nytt lovbrudd eller ikke, og diskuterer hvorvidt den er diskriminerende eller ikke. Innledningsvis presenterer vi fordelingen av prediksjoner gjort av algoritmer over ulike grupper som kvinner, menn, hvite, og afroamerikanere og diverse partial dependency plots. Fra disse resultatene observerte vi at afroamerikanere som gruppe feilpredikeres oftere enn hvite i tillegg til at tidligere forbrytelser hos afroamerikanere hadde større bestemmelse for om algoritmen predikerte at de begikk nye lovbrudd enn hos hvite. Begge observasjoner motiverer en nærmere diskusjon av spørsmålet om hvorvidt algoritmen er rettferdig eller ikke. I diskusjonen presenterer vi seks ulike definisjoner av rettferdighet og undersøker hvorvidt COMPAS-algoritmen etterfølger hver av disse definisjonene. Vi diskuterer videre om det er mulig for COMPAS-algoritmen å etterfølge alle de seks definisjonene og vurderer noen potensielle avveininger som må bedømmes i denne sammenheng.

Introduksjon

Denne rapporten består av en analyse av COMPAS-algoritmen, et beslutningsverktøy som er blitt brukt til å bestemme om en tidligere innsatt kommer til å begå et nytt lovbrudd eller ikke. Målet med rapporten er å finne ut om Probublica sin kritikk av algoritmen stemmer og å belyse noen relevante hensyn som må tas for å bestemme algoritmisk rettferdighet. Vi vil aller først legge frem resultater som kan tyde på at COMPAS-algoritmen er diskriminerende og som ligner på de resultatene som Probublica tok som utgangspunkt i deres endelige konklusjon om at COMPAS-algoritmen er diskriminerende. Videre diskuterer vi mer detaljert hvordan man kan vurdere om COMPAS-algoritmen er rettferdig eller ikke. I lys av det som kommer frem i diskusjonen mener vi denne rapporten er et godt eksempel på det Kleinberg et al. (2019, s.1) mener er fordelaktig ved bruk av algoritmer i beslutningsprossesser. Innsikt i deres beslutningstagning kan lettere la seg analysere. Dermed er det også lettere å avdekke diskriminering når en algoritme er involvert i beslutningsprossesen.

Metode

Vi har brukt en rekke forskjellige metoder for å komme frem til resultatene som ligner på de Propublica kom frem til og til de resultatene som vi gjør bruk av i diskusjonsdelen. Under følger en oversikt over hvordan vi har produsert disse resultatene.

Sortering av data

For å se på hvordan COMPAS-algoritmen vurderer ulike identitetsgrupper, ser vi på distribusjonen av risikoscorene fra COMPAS blant de forskjellige gruppene. Først grupperer vi dataene etter kjønn og etnisk bakgrunn (kvinner og menn, hvite og afroamerikanere). Med utgangspunkt i disse gruppene, utforsker og filtrerer vi dataene ytterligere for å ekstrahere følgende informasjon:

1. Andel av hver gruppe som ble predikert til å gjenta forbrytelser i framtiden, versus andel som faktisk gjorde det. Vi valgte å behandle kandidater med risikoscorene "Medium" og "High" som positiver; altså at COMPAS predikerer at disse kandidatene kommer til å begå flere lovbrudd. Scoren “Low” anser vi som en negativ prediksjon. Vi sammenlikner risikoscorene til kandidatene med hvorvidt de begikk et nytt lovbrudd innen to år.

2. Hvor mange av kandidatene innenfor hver gruppe som fikk de ulike risikoscorene.
3. Feilprediksjoner for svarte og hvite. Dette innebærer å finne antall falske negativer og falske positiver for hver av gruppene. Vi normaliserer disse målene i forhold til det totale antallet kandidater innenfor sine respektive grupper (afroamerikanere/ hvite). Grunnen til dette er for å ta hensyn til at det er ulikt antall afroamerikanere og hvite i datasettet.

Dette svarer til henholdsvis figur 1, 2 og 3 nedenfor.

Fremstilling av Forvirringsmatrise

For å få en oversiktlig og informativ fremstilling av COMPAS sine prediksjoner, og sette dem i sammenheng med det faktiske utfallet, lagde vi forvirringsmatriser for gruppene "hvite", "afroamerikanere", "kvinner" og "menn", samt hele populasjonen. For å gjøre dette måtte vi først dele opp datasettet i undergrupper med utvalget nevnt ovenfor. Deretter kjørte vi hver undergruppe gjennom en prosess som sjekker hva COMPAS predikerte for hver person i undergruppen opp mot det faktiske resultatet for den personen. Dette ga ett av fire mulige svar. Enten predikerte COMPAS at personen:

1. ville begå et nytt lovbrudd innen to år, og fikk *rett* i det (altså begikk personen faktisk et nytt lovbrudd). Dette kalles da en "sann positiv".
2. ville begå et nytt lovbrudd innen to år, og tok *feil*. Dette kalles en "falsk positiv".
3. *ikke* ville begå et nytt lovbrudd innen to år, og fikk *rett* i det. Dette kalles en "sann negativ".
4. *ikke* ville begå et nytt lovbrudd innen to år, og tok *feil* av det. Dette kalles en "falsk negativ".

Vi kan altså se at knaggene "positiv"/"negativ" følger COMPAS sin prediksjon av om personen vil begå et nytt lovbrudd, og knaggene "sann"/"falsk" henviser til hvorvidt prediksjonen var i samsvar med det faktiske resultat.

Da vi kjørte undergruppen gjennom prosessen som sjekket COMPAS sin prediksjon opp mot det faktiske resultatet for hver person, fikk vi det totale antallet av sanne positive, falske positive, sanne negative og falske negative for hver gruppe. Dette lagret vi i hver gruppes forvirringsmatrise.

Hver gruppes forvirringsmatrise fikk så en tilhørende tabell med ulike ytelsesindekser beregnet. Disse indeksene er hentet fra artikkelen "Fairness in Criminal Justice Risk Assessments: The State of the Art" (Berka et al., 2017), og vil bli lagt frem og diskutert senere i rapporten.

Til slutt ble forvirringsmatrisene og den tilhørende tabellen med ytelsesindekser visualisert.

Trening av modeller

For å kunne utføre ulike sjekker på datasettet, slik som å se etter bias i COMPAS sine prediksjoner ved hjelp av Partial dependence, trente vi en logistisk regresjonsmodell. Modellen skulle predikere hvorvidt en person ville bli vurdert til høy eller lav risikoscore av COMPAS-algoritmen. For å gjøre dette trente vi den på det samme datasettet, men fjernet alle datapunkter knyttet til COMPAS modellens risikoscore. Modellens treffsikkerhet på rett under 75% henviser dermed til hvor ofte den predikerte COMPAS sin risikoscore for personen korrekt.

Fremstilling av partial dependence

For å vite hvilken input som er mest avgjørende for prediksjonene til COMPAS, ønsker vi også å kartlegge hver parameters innvirkning på beslutningen til modellen, når alle de andre parametrene holdes konstant. Er det for eksempel slik at et høyt antall tidligere begåtte lovbrudd vil gjøre modellen mer tilbøyelig til å predikere nye lovbrudd i fremtiden? Strategien for å finne hvordan modellens output er delvis avhengig av hver parameter, kalles *partial dependence*. Vi plotter parametrene som omhandler etnisitet for seg og de øvrige parametrene for seg. Resultatet er partial dependency-plots (fremover omtalt som PDP-er) som illustrerer hvordan konfigurasjonen til hver parametrene påvirker modellens beslutning i ulike retninger og omfang.

For å oppdage eventuell forskjellsbehandling, ønsker vi også å undersøke hvorvidt ulike parametre gjør seg mer eller mindre gjeldende for svarte versus hvite. Med andre ord; hva ligger til grunn for at en svart kandidat predikeres til å begå et nytt lovbrudd? Skiller dette seg fra årsakene til den tilsvarende prediksjonen for en hvit kandidat? For å finne ut hvilke parametre som er mest avgjørende for henholdsvis svarte og hvite, deler vi datasettet i to - ett med kun hvite og ett med kun svarte - og trener én modell på hver av dem. Vi genererer så PDP-er som illustrerer *partial dependence* for hver av disse modellene, for hver parameter.

Nøytrale modeller vi kjennetegnes ved marginale avvik i hver parameters delaktige avhengigheten, for hvite versus svarte. Store avvik mellom *dependency*-linjene i et PDP forteller oss at den aktuelle parameteren vektlegges i ulik grad for henholdsvis hvite og svarte amerikanere.

Idet vi vurderer og sammenlikner *partial dependence* for hver parameter for henholdsvis hvite og svarte, gjør vi en antakelse om at datasettets størrelse har liten eller ingen innvirkning på resultatene. Det vil si; vi antar at modellene våre er trent på tilstrekkelige mengder data om hvite og svarte, slik at eventuelle forskjeller i *partial dependence* skyldes én eller annen form for bias.

Resultater

Her presenterer vi noen resultater som viser hvordan COMPAS algoritmens prediksjoner fordeler seg på ulike grupper og hvilke parametere som er bestemmende for hvorvidt medlemmer av en spesifikk identitetsgruppe (f.eks alle hvite) predikeres til å begå nye lovbrudd eller ikke.

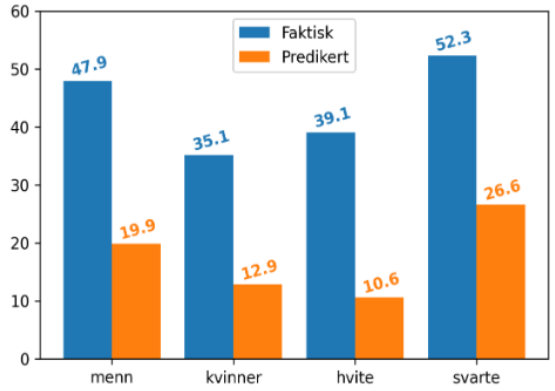
Utforsking av data

Figur 1 viser at hvite blir predikert til å gjenta forbrytelser sjeldnest relativt sett ($\frac{10.6\ \%}{39.1\ \%} \approx 0.27$). Afroamerikanere blir predikert til å gjenta forbrytelser om lag halvparten så ofte ($\frac{26.6\ \%}{52.3\ \%} \approx 0.51$) som de faktisk gjør det, hvilket er det høyeste blant gruppene.

Figur 2 illustrerer at det er klart flere menn enn kvinner, og flere svarte enn hvite, i datasettet. Den viser også at det er en forholdsvis jevn fordeling av risikoscorene blant svarte, mens de fleste som er hvite får en lav risikoscore. Gitt at vi behandler medium og høy risikoscore som det å bli predikert til å begå nye lovbrudd, tildeles en større andel afroamerikanere denne klassifikasjonen enn hvite.

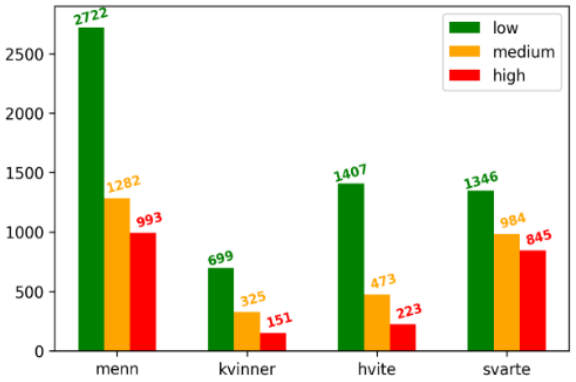
I figur 3 kan vi se at svarte er den gruppen som oftest predikeres til å begå lovbrudd når de ikke faktisk gjør det (20,19 %), mens hvite blir oftest predikert til å ikke begå lovbrudd når de faktisk gjør det (19,4 %).

Andel av kandidatene som predikeres til å gjenta forbrytelser i framtiden, versus andel av kandidatene som faktisk gjorde det.



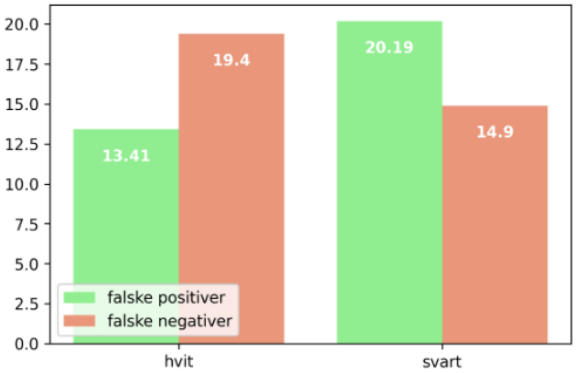
Figur 1

Hvor mange av kandidatene innenfor hver gruppe som fikk ulike risikoskårer *



Figur 2

Feilprediksjoner for svarte og hvite (andel).



Figur 3

Forvirringsmatrise for Afroamerikanere

	Sanne	Falske
Positive	1188	641
Negative	873	473

Figur 4

Forvirringsmatrise for Hvite

	Sanne	Falske
Positive	414	282
Negative	999	408

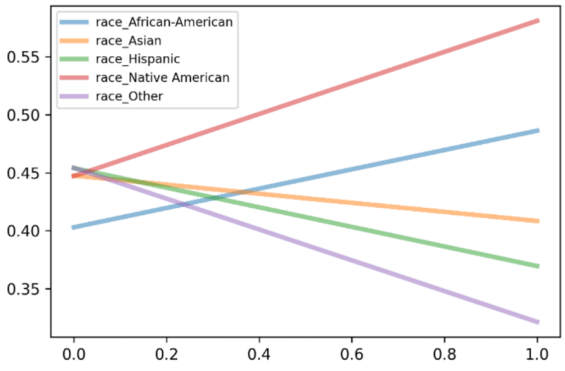
Figur 5

Forvirringsmatriser

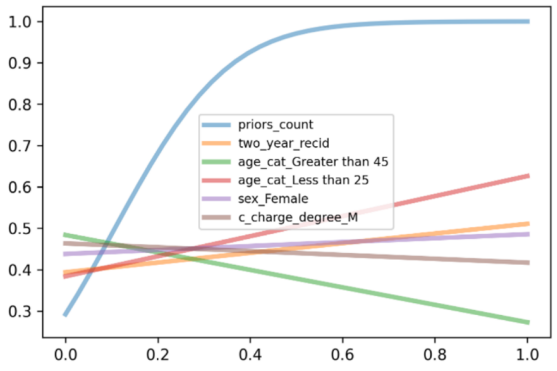
I figur 4 og 5 kan vi se forvirringsmatriser for afroamerikanere og hvite i datasettet. Dette er en type datafremstilling som skal vise antall riktige og gale prediksjoner gjort av COMPAS-algoritmen, målt mot det faktiske utfallet. Her er det noen punkter som er verdt å merke seg. Første punkt er at hvite oftere blir predikert negativt av COMPAS enn afroamerikanere, altså predikerer algoritmen oftere at hvite ikke vil gjenta en forbrytelse innen to år enn svarte. Særlig merkverdig er dette fordi gruppen av hvite er betydelig mindre enn gruppen av svarte, dermed er det en disproporsjonal forskjell i antall negative prediksjoner mellom gruppene. Tilsvarende kan vi se at det motsatte holder: det er et disproporsjonalt større antall afroamerikanere som predikeres til å gjenta forbrytelse innen to år etter utslipp.

Partial dependency

Foruten urfolk[1], er *dependency*-linjen for "race African-American" den eneste etnisitetbaserte parameteren med positivt stigningstall. Det å være afroamerikaner styrker altså modellens beslutning om at kandidaten vil begå nye lovbrudd.



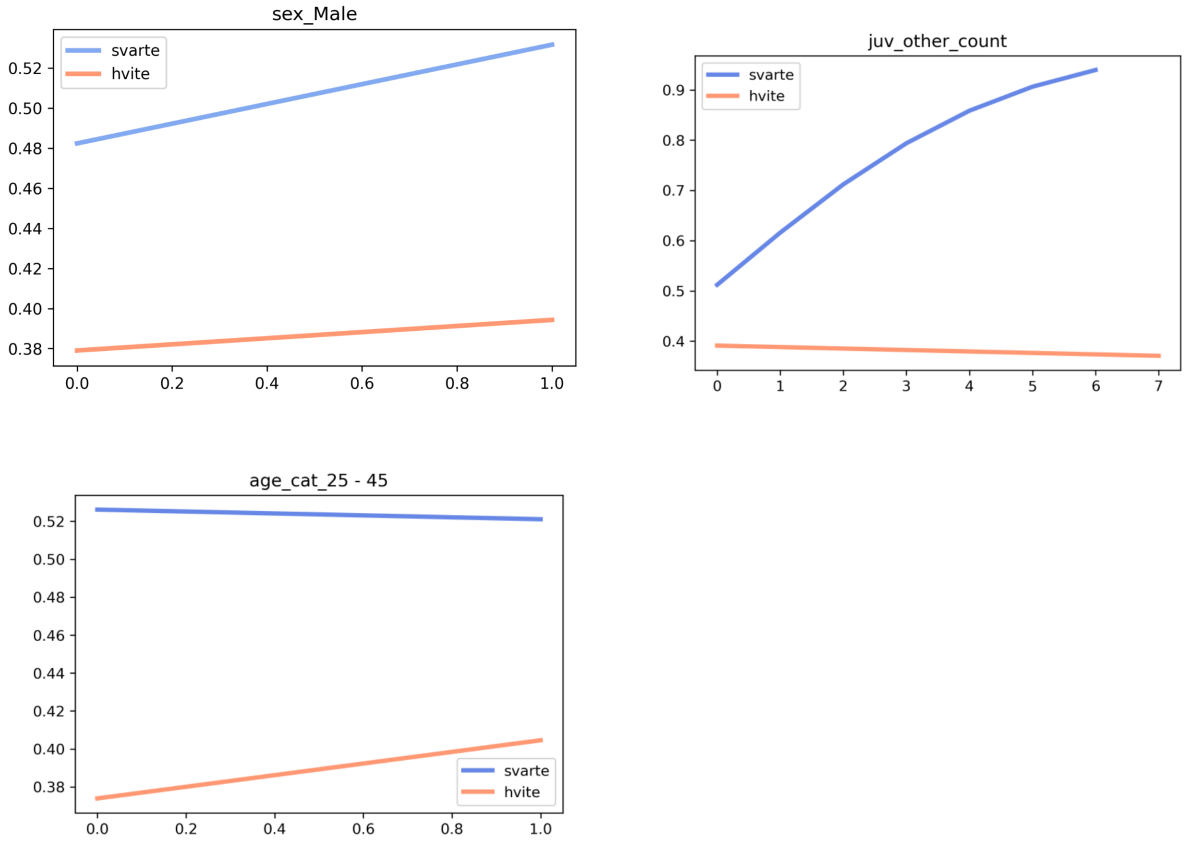
Figur 8: *partial dependence* for de etnisitetsbaserte parametrene i Propublica-modellen.



Figur 9: *partial dependence* for de øvrige parametrene i Propublica-modellen.

Figur 6

[1] Det er kun 11 urfolk (0.17 % av datasettet), så denne *dependency*-linjen er lite informativ.



Figur 7

Hvite "straffes" for å være mellom 25 og 45 år gamle (beslutningen om gjentakende lovbrudd styrkes), mens det motsatte er tilfellet for afroamerikanere i samme aldersgruppe. Å være en svart mann (i motsetning til det å være svart kvinne) straffes derimot mer enn det tilsvarende for hvite. Et høyt antall kriminelle handlinger i ung alder (foruten forbrytelser og forseelser) styrker modellens sikkerhet på at en svart kandidat vil gjenta forbrytelser betraktelig; det motsatte er tilfellet for hvite med det tilsvarende antallet kriminelle handlinger.

Diskusjon

Resultatene over kan indikere at COMPAS-algoritmen er diskriminerende mot afroamerikanere. For å undersøke dette nærmere har vi valgt å vurdere hvor rettferdig COMPAS-algoritmen er ved å sammenligne dens vurdering av to grupper: hvite og afroamerikanere. For å vurdere om COMPAS er rettferdig eller ikke presenterer vi seks ulike definisjoner av rettferdighet og analyserer hvor godt COMPAS scorer på hver av disse forståelsene av rettferdighet.

Seks ulike måter å definere algoritmisk rettferdighet på

Vi har valgt å bruke de seks ulike definisjonene av algoritmisk rettferdighet fra Berka et al. Disse definisjonene tar utgangspunkt i fordelinger av riktige- og feilprediksjoner blant grupper. Dermed er informasjonen fra forvirringsmatrisene essensiell for å vurdere om COMPAS er rettferdig ut i fra disse seks definisjonene.

a := sanne positive

b:= falske negative

c := falske positive

d: = sanne negative

1. **Overall Accuracy Equality:** Algoritmens prediksjons suksess skal være lik blant alle grupper. Antall sanne positive og negative skal derfor være lik blant gruppene, eller at (a+d)/(a+b+c+d) skal være lik for alle grupper.
2. **Statistical parity:** Andelen av positive skal være lik for hver gruppe, dette medfører nødvendigvis også at andelen negative blir lik for hver gruppe. Dette er ekvivalent med at (a+c)/(a+b+c+d) skal være lik for alle grupper
3. **Conditional Procedure Accuracy Equality:** Kostnaden av en sann positiv sett i forhold til falske negative og kostnaden av en sann negativ sett i forhold til en falsk positiv skal være lik blant alle grupper. Dette er ekvivalent med at a/(a+b) og d/(c+d) skal være lik for alle grupper
4. **Conditional Use Accuracy Equality:** Andelen sanne positive av de totale predikerte positive og andelen av sanne negative av de totale predikerte negative er lik for alle grupper. Dette er ekvivalent med at a/(a+c) og d/(d+b) skal være lik for alle grupper
5. **Treatment Equality:** Ratioen av feilprediksjoner skal være lik for alle grupper. Dette er ekvivalent med at b/c eller c/b skal være lik for alle grupper.
6. **Total Fairness:** Denne definisjonen inkorporerer alle de tidligere 5. Det vil si at alle de 5 forhold er oppfylt.

COMPAS i forhold til de seks rettferdighetsdefinisjonene

Vi skal nå analysere hvordan COMPAS-algoritmen legger seg i forhold til alle disse seks kriteriene. Under følger det en oversikt der vi har regnet ut indeksene som er gitt av de 6 rettferdighetsdefinisjonene over med tall fra forvirringsmatrisene for hvite og afroamerikanere fra analysen av COMPAS-algoritmen.

PPV := Positiv prediktiv verdi (a / (a+c))

NPV := Negativ prediktiv verdi (d / (d+b))

SPP := Sanne prediksjoner delt på populasjonen ((a+d)/(a+b+c+d))

PPP := Positive prediksjoner for populasjonen ((a+c)/(a+b+c+d))

SPFN := Sanne positive delt på sanne positive og falske negative ((a)/(a+b))

SNFP := Sanne negative delt på sanne negative og falske positive ((d)/(c+d))

FPFN := Falske positive per falske negative (c/b)

Forhold for forvirringsmatrise for Hvite

	PPV	NPV	SPP	PPP	SPFN	SNFP	FPFN
Verdi	0.59	0.71	0.67	0.33	0.5	0.78	0.69

Forhold for forvirringsmatrise for Afroamerikanere

	PPV	NPV	SPP	PPP	SPFN	SNFP	FPFN
Verdi	0.65	0.65	0.65	0.58	0.72	0.58	1.36

Fra oversikten er det tydelig at COMPAS-algoritmen kan forstås som rettferdig ut ifra definisjon 1 og definisjon 4 siden SPP er tilnærmet lik for begge grupper og fordi PPV og NPV parvis også er ganske nærme hverandre. Algoritmen har et betydelig avvik mellom henholdsvis SPFN og SNFP mellom de to gruppene og dermed er den langt unna å overholde kriteriene fra definisjon 3. De parametrene der det er svært store avvik mellom hvite og afroamerikanere er PPP og FPFN som er indeksene for henholdsvis definisjon 2 og 5. Dermed er COMPAS-algoritmen svært lite rettferdig i henhold til 2 og 5, som også gjør den lite rettferdig ut i fra definisjon 6 som krever at alle de andre fem definisjonene overholdes samtidig. Resultatene er oppsummert i tabellen under.

Rettferdighetsdefinisjon	**I hvor høy grad COMPAS overholder kriteriene fra definisjonen**
1	Overholdes i høy grad
2	Overholdes i liten grad
3	Overholdes i middels grad
4	Overholdes i høy grad
5	Overholdes i liten grad
6	Overholdes i liten grad

Altså kan COMPAS-algoritmen forstås som rettferdig eller urettferdig avhengig av hvilken definisjon man vil bruke. Spørsmålet er om COMPAS-algoritmen kan forbedres slik at den etterfølger alle definisjonene. Med andre ord, er det mulig å modifisere COMPAS-algoritmen slik at den etterfølger definisjon 6?

Et umulighetsteorem

Resultatet fra undersøkelsen over, at definisjon 4 og 1 overholdes, men 5, 2 og dermed også 6 ikke overholdes, kan forklares og til en hvis grad generaliseres i et teorem. Vi presenterer to definisjoner som inngår i teoremet først.

Base rate: Hyppigheten av et fenomen blant en gruppe. I denne sammenheng vil det være relevant å snakke om base raten til afroamerikanere som begår nye lovbrudd.

Separasjon: Et datasett er separabelt dersom det finnes en predikator som kan klassifisere medlemskap i alle grupper med 100% sikkerhet.

Teorem: Når ulike grupper har ulik base rate og når det ikke er noen separasjon i datasettet kan ikke rettferdighetsdefinisjon 4 være oppfylt og at det samtidig er lik rate av falske negative og falske positive blant alle gruppene.

Bevis: Vi refererer til Berka et al 2017 s. 17. for henvisning til et bevis.

□

Teoremet over viser at oppfyllelsen av definisjon 4 er inkompatibel med potensielt flere av de andre definisjonene når det er ulik base ratene blant gruppene og når det er null separasjon. Særlig er definisjon 5 vanskelig å oppfylle samtidig som 4 fordi definisjon 5 bestemmes av kun falske positive og falske negative som umulig kan være like for alle grupper gitt at definisjon 4 er oppfylt.

Vi antar at datasettet som COMPAS anvendes på ikke er separabelt. Hovedbegrunnelsen for dette er at de fleste empiriske datasett ikke er separable (Berka et. al. 2017, s. 18-19) Videre er det klart at det er ulik base rate mellom svarte og hvite (se "Resultater"-seksjonen). Dermed gjelder umulighetsteoremet for COMPAS-algoritmen som kan forklare hvorfor algoritmen overholt rettferdighetsdefinisjon 4, men ikke 2 og 5. COMPAS-algoritmen kan derfor umulig oppfylle alle de 6 kriteriene samtidig, og dermed må man avveie hvilket kriterium som er mest relevant for å bestemme algoritmens rettferdighet. Rettferdighet blir i COMPAS-algoritmens tilfelle, og alle andre tilfeller hvor base raten mellom flere grupper er forskjellige, relativt til hvilken rettferdighetsdefinisjon som man tar utgangspunkt i. Å konkludere med at COMPAS-algoritmen er urettferdig fullt og helt, slik det kan synes at Propublica har gjort, er dermed en forhastet og lite nyansert konklusjon. COMPAS algoritmen er tross alt rettferdig etter definisjon 2 og 4, to definisjoner som isolert sett virker svært rimelige.

Hvordan kan COMPAS bli mer rettferdig, hvilke avveininger bør tas?

Umulighetsteoremet over viser at det finnes en avveiningen med tanke på COMPAS-algoritmens rettferdighet: Bør raten av feilprediksjoner dvs., både falske positive og falske negative være lik for alle grupper eller bør definisjon 4. være det relevante rettferdighetsprinsipp å følge?

Konsekvensene av å velge definisjon 4 som styrende rettferdighetsdefinisjon er at fordelingen av feilprediksjoner blir ulik blant identitetsgruppene. En slik fordeling kan virke urettferdig dersom antall falske negative er svært høy for en gruppe mens antall falske positive er høy for en annen gruppe og dersom falske positive blir ansett som en mye mer kostbar feilprediksjon en falske negative.

Konsekvensene av å velge definisjon 5, eller mer generelt, å kreve at fordelingen av falske positive og negative skal være lik for alle grupper er at det kan resultere i en dårligere prediktiv ytelse. Med andre ord, PPV og NPV vil variere mellom grupper.

I lys av dette blir avveiningen som må tas følgende:
Enten:

(a) Fordelingen av prediktiv suksess (antall sanne positive og negative) er lik blant alle grupper .

Eller:

(b) Fordelingen av falske positive og negative er lik blant alle grupper.

Dersom man velger å etterfølge (b) går dette utover algoritmens evne til å predikere riktig siden etterfølgelsen av (b) direkte påvirker etterfølgelsen av (a) negativt. Dermed blir algoritmen mindre aktuell for bruk i beslutningsprossesser. Likefullt kan (b) være det riktige valget i en sammenheng hvor falske positive blir ansett som ekstremt kostbare. I COMPAS-algoritmens tilfelle er prediksjonsevne viktig i tillegg til at kostnaden ved en falsk positiv er høy. Derfor er det svært vanskelig å bestemme hvorvidt man burde velge (a) eller (b) som styrende for hvilken rettferdighetsdefinisjon algoritmen burde etterfølge. Vi vil ikke ta en slik avveining her. Derimot vil vil igjen understreke at denne avveiningen viser at fullkommen rettferdighet i COMPAS-algoritmens tilfelle ikke er mulig. Dermed må en konklusjon angående algoritmens rettferdighet være relativ til hvilken definisjon av rettferdighet tar utgangspunkt i. Dette hensynet mangler i Probublica sin rapport.

Konklusjon

Gjennom vår undersøkelse av COMPAS-algoritmen har vi avdekket at svaret på spørsmålet knyttet til algoritmens rettferdighet avhenger av hvilken forståelse av rettferdighet som er lagt til grunn. Under visse definisjoner på rettferdighet kan COMPAS-algoritmen ansees som rettferdig; under andre kan den ikke det. Videre har vi vist at COMPAS-algoritmen umulig kan modifiseres til å etterfølge alle former for rettferdighet samtidig og at avveining mellom hvilken definisjon av rettferdighet som bør etterfølges er opp til interessentene. Vi mener at vår undersøkelse av COMPAS-algoritmen styrker hypotesen til Kleinberg et. al. 2019 (s. 1)

når det gjelder algoritmer versus mennesker i beslutningsprosesser. Grunnen er at vi ikke ser for oss at en tilsvarende analyse av beslutningsprossesen til et menneske kunne blitt gjort. Dersom det kun hadde eksistert data fra menneskelig beslutningtagning om hvorvidt tidligere innsatte begår nye lovbrudd hadde det dermed vært vanskeligere å avdekke de nyansene og avveiningene som gjelder når en skal vurdere om en beslutning er rettferdig eller ikke i et slikt tilfelle. Derfor burde algoritmer brukes i beslutningsprosseser som et verdifullt analyseverktøy for å avsløre de avveiningene som må tas for å vurdere om en beslutning er rettferdig eller ikke.

Litteratur

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R. Sunstein. (2019). "Discrimination in the Age of Algorithms" i *National Bureau of Economic Research,* Cambridge MA.

Richard Berka,b, Hoda Heidaric , Shahin Jabbaric , Michael Kearnsc , Aaron Rothc. (2017). "Fairness in Criminal Justice Risk Assessments: The State of the Art"
DOI:
https://doi.org/10.1177/0049124118782533

Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner. (2016). "Machine Bias" i *Propublica*
URL:
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing