Analyse av COMPAS

14.05.2024

Bjarte Holgersen **Cecilie Sivertsen Rønnestad** **Sadaf Samin Dar**
Simon Lyng-Jørgensen **Åsmund Olai Sand-Larsen**

Innhold

1. 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.

1.1. 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.

2. Teori

2.1. 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.

2.2. 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).

2.3. 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).

2.4. 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.

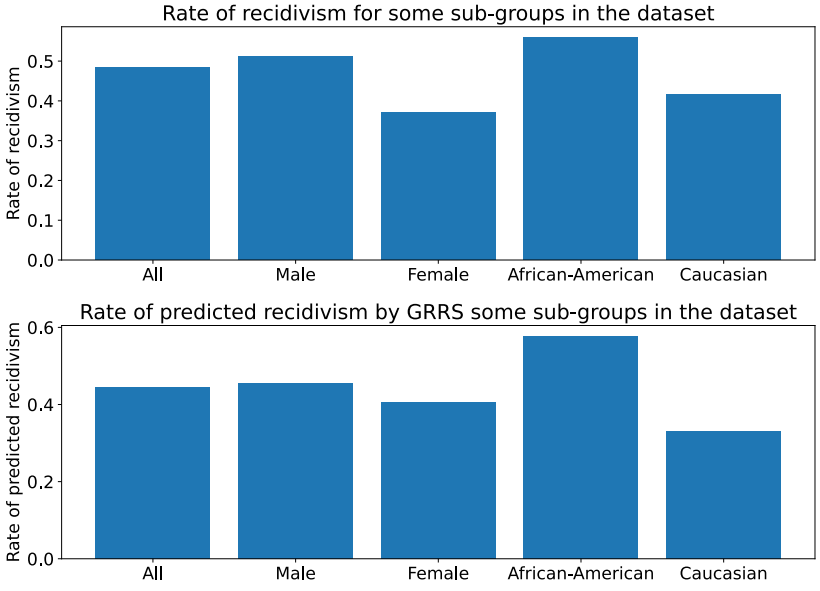
2.5. Skjevhet fra Bayes' setning

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:

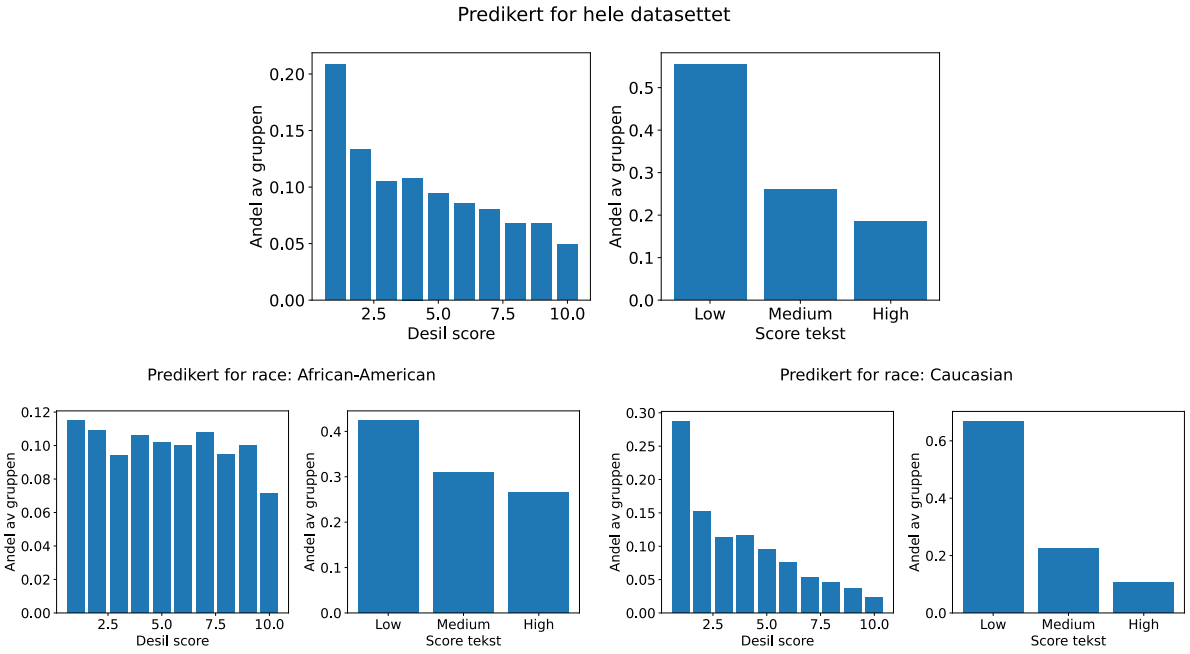
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

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.

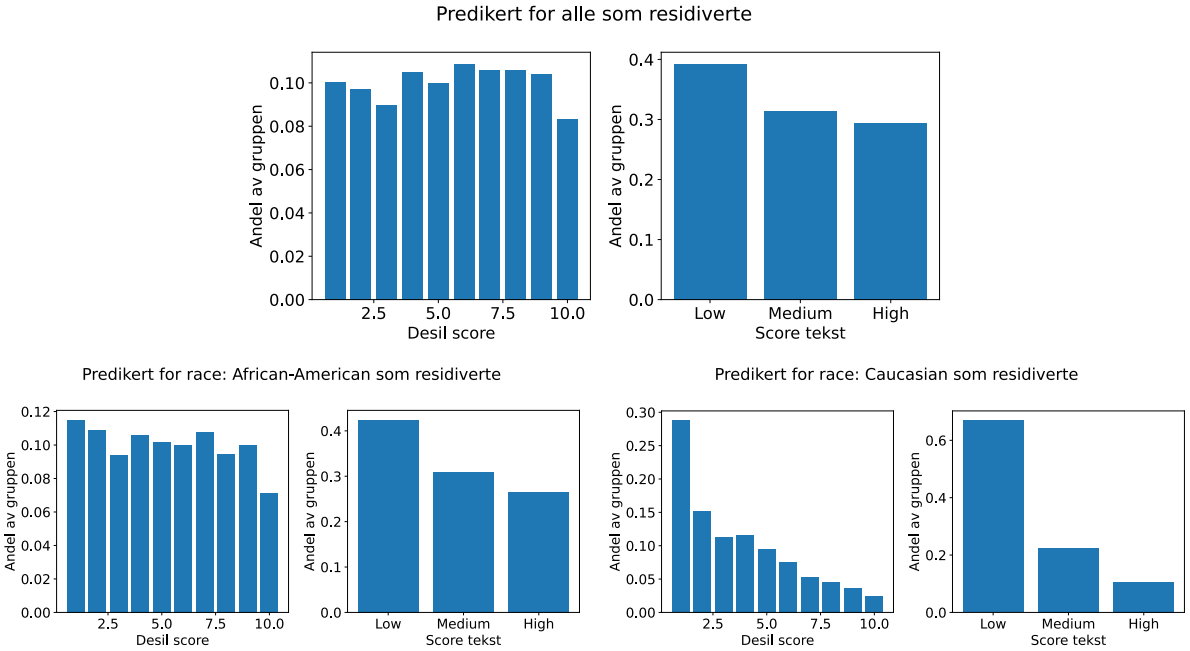
3. Resultater



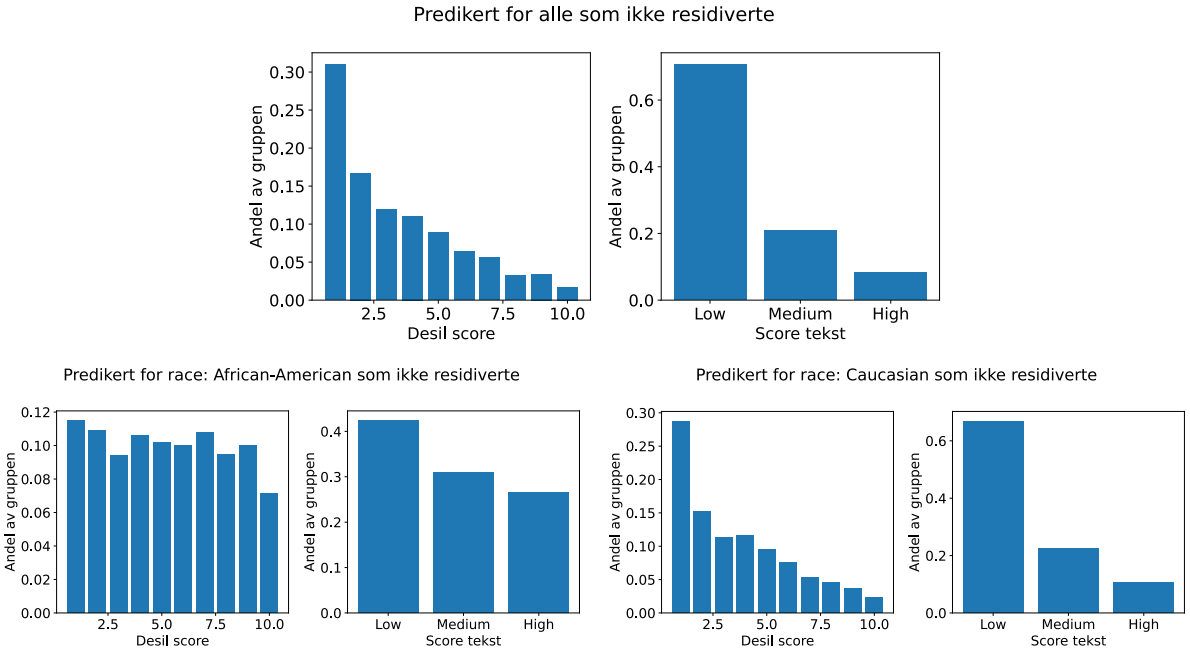
Figur 1: Residivismerater for noen undergrupper i datasettet (øverst) og raten av predikert residivisme fra GRRS (nederst). Det er antatt at en ikke-lav risiko-score (tilsvarer desil 5-10) er predikert til å residivere.



Figur 2: Det som ble predikert av GRRS for hele datasettet (øverst), afroamerikanere (venstre) og hvite (høyre).



Figur 3: Det som ble predikert av GRRS, gitt residivisme, for hele datasettet (øverst), afroamerikanere (venstre) og hvite (høyre).



Figur 4: Det som ble predikert av GRRS, gitt ingen residivisme, for hele datasettet (øverst), afroamerikanere (venstre) og hvite (høyre).

Tabell 1: Sannhetstabell for hele datasettet. Det er her antatt at en ikke-lav prediksjonstekst (tilsvarer desi-score 5-10) er at algoritmen predikerer residivisme, og sannhetstabellen er deretter beregnet som vanlig ved binær klassifisering.

All	Actual true	Actual false
Predicted true	0.29	0.15
Predicted false	0.19	0.36

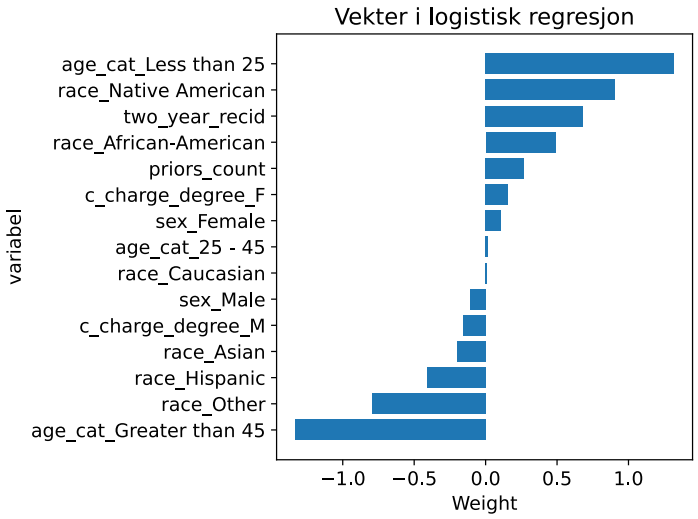
Tabell 2: Sannhetstabell for undergruppene afroamerikanere (oppe til venstre), hvite (oppe til høyre), kvinner (nede til venstre) og menn (nede til høyre). Som i Tabell 1, er det antatt at en ikke-lav prediksjonstekst er predikert residivisme og sannhetstabellene er deretter beregnet som vanlig ved binær klassifisering.

race: African-American	Actual true	Actual false
Predicted true	0.39	0.18
Predicted false	0.17	0.26

race: Caucasian	Actual true	Actual false
Predicted true	0.20	0.13
Predicted false	0.21	0.46

sex: Female	Actual true	Actual false
Predicted true	0.22	0.19
Predicted false	0.15	0.44

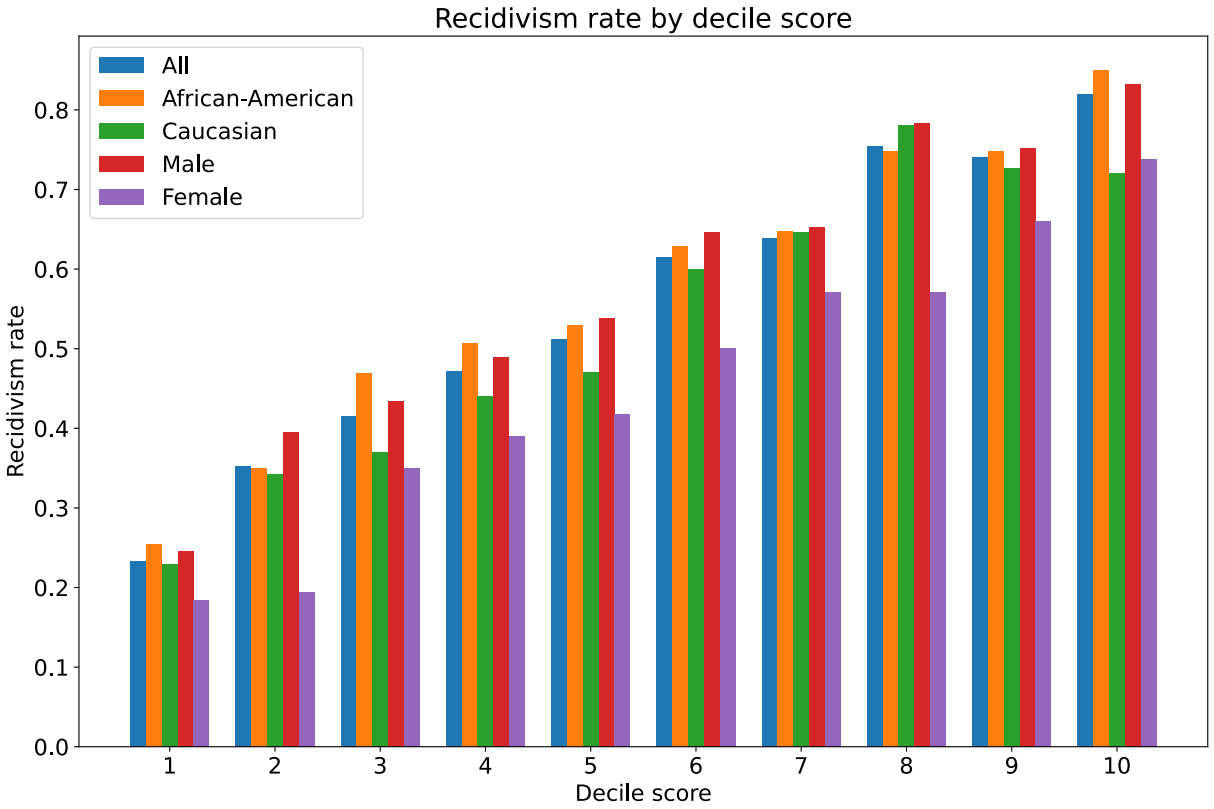
sex: Male	Actual true	Actual false
Predicted true	0.31	0.14
Predicted false	0.20	0.35



Figur 5:

Tabell 3: Raten for residivisme for de forskjellige risiko-scorene. Altså er hvert tall her residivisme-raten for de i datasettet som er tilhører gruppen (kolonne) og som fikk scoren av GRRS (raden). Relativ forskjell mellom **African-American** (a) og **Caucasian** (c) er beregnet ved $\frac{a-c}{c}$.

Risiko-score	All	African-American	Caucasian	Relativ forskjell
Low	0.343	0.390	0.316	0.23602
Medium	0.585	0.603	0.556	0.08384
High	0.767	0.775	0.749	0.03508
1	0.233	0.255	0.230	0.10900
2	0.353	0.350	0.343	0.02052
3	0.416	0.470	0.370	0.27059
4	0.471	0.507	0.440	0.15236
5	0.512	0.529	0.470	0.12641
6	0.614	0.629	0.600	0.04822
7	0.639	0.647	0.646	0.00188
8	0.755	0.748	0.781	−0.04319
9	0.740	0.748	0.727	0.02800
10	0.819	0.850	0.720	0.18086



Figur 6: Residivisme-rater ved de forskjellige desil-scorene, samme data som i Tabell 3 men her også med kvinner og menn.

4. Diskusjon

4.1. Våre funn

4.2. Northpointe vs ProPublica

ProPublica har publisert en studie som hevder at COMPAS diskriminerer afro-amerikanere. Northpointe på sin side har kommet med et tilsvar som hevder at COMPAS ikke diskriminerer. For å forstå kjernen av uenigheten mellom Northpointe og ProPublica er det først viktig å forstå hva begge partene hevder, på hvilke punkter er de enige og på hvilke punkter er de uenige.

ProPublica fant at algoritmen predikerer sannsynligheten for å «reoffend», altså gjøre en kriminell handling på nytt når du tidligere har blitt arrestert for en kriminell handling, til å være like nøyaktig for både afro-amerikanere og hvite kriminelle. Sannsynligheten for at algoritmen predikerte riktig var 66 % for afro-amerikanere og 59 % for hvite. Dette mener Northpointe er et viktig poeng, fordi det viser en likestilling i nøyaktigheten av modellen. Uenigheten ligger i at ProPublica hevder at det er store forskjeller i algoritmen i de tilfellene den tar *feil*.

ProPublica hevder at når algoritmen tok feil om «reoffending» på en skjev måte, ved å predikere at afro-amerikanere ville gjenta kriminalitet når de endte opp med å ikke gjøre det, dobbelt så mye for afro-amerikanere sammenlignet med hvite. Dette kalles en falsk positiv feil, altså at modellen finner et positivt funn (sannsynligheten for å begå en kriminell handling på nytt) når det faktisk ikke skjer. ProPublica hevder at de også kontrollerte for tidligere kriminell historikk, alder og kjønn, men fortsatt hadde afro-amerikanere 77 % høyere sannsynlighet for å bli klassifisert som en med høy risiko for å begå kriminelle, voldelige handlinger, og 45 % mer sannsynlighet for å bli klassifisert som en som skal begå kriminelle handlinger. Men den faktiske sannsynligheten for å begå kriminelle handlinger viser

små forskjeller mellom afro-amerikanere og hvite, med henholdsvis 63 % sannsynlighet for å afro-amerikanere, og 59 % for hvite.

Også omvendt hevder ProPublica at algoritmen tok feil på en skjev måte, ved å si at man hadde lav sannsynlighet for å begå kriminalitet på nytt når man ender opp med å faktisk gjøre det, med henholdsvis 48 % for hvite og 28 % for afro-amerikanere. For kriminelle, voldelige handlinger var det enda større forskjeller med at hvite hadde 63 % større sannsynlighet for å bli klassifisert som lav-risiko når de endte opp med å faktisk begå kriminelle, voldelige handlinger. Dette kalles en falsk negativ feil, altså at modellen finner et negativt funn (sannsynligheten for å ikke begå en kriminell handling) når det egentlig burde være et positivt funn.

Northpointe hevder på sin side at ProPublica tar feil, fordi deres statistiske metode ikke har tatt høyde for at det er forskjellig "base rates of recidivism" blant afro-amerikanere og hvite. «Recidivism» viser til å bli arrestert for kriminalitet på nytt når man har blitt arrestert for kriminelle handlinger tidligere. Her mener Northpointe at ProPublica burde ha inkludert deskriptiv statistikk som viser at afro-amerikanere har høyere risiko for å begå kriminelle handlinger på nytt, på grunn av relaterte risikofaktorer slik som narkotika, alder, alder ved første arrest, antall tidligere arrest, og utdanning. Særlig alder mener Northpointe er en sterk predikator for å begå kriminelle handlinger på nytt, ved at unge har høyere risiko. Det hvite utvalget som var en del av utvalget til ProPublica var mye eldre enn den svarte befolkningen. ProPublica har også selv skrevet at alder var en sterkere prediktor enn rase på risikoen for å begå kriminelle handlinger på nytt.

Videre mener Northpointe at ProPublica også tolker feil når de skal snakke om hva slags feil algoritmen tar, hvor de skiller mellom "model errors" og "target population errors". «Model errors» henger sammen med sensitiviteten og spesifisiteten til en modell. Sensitiviteten til en modell, også kalt ekte positiv, er andelen gjentagende-kriminelle som faktisk blir klassifisert som gjentagende-kriminelle. Det komplementære til dette falske negative, altså gjentagende-kriminelle som falskt blir klassifisert som ikke-gjentagende kriminelle. Spesifisiteten til en modell, også kalt falsk negativ, er andelen ikke-gjentagende kriminelle som faktisk blir klassifisert som dette. Det komplementære til dette er falsk positive, altså ikke-gjentagende kriminelle som blir klassifisert som gjentagende-kriminelle. Det var nettopp «model errors» som ProPublica hevder viser en skjevhet mot afroamerikanere. Northpointe hevder at «model errors» ikke er et valid mål for å se om det er skjevhet i dataen, fordi for det første kan ikke en dommer bruke sensitiviteten eller spesifisiteten på individnivå fordi på det tidspunktet man tar i bruk modellen vet man ikke om den personen vil gjenta kriminalitet eller ikke. Videre poengterer Northpointe at det er urealistisk å forvente at sensitiviteten og spesifisiteten vil være helt lik i to utvalg som har forskjellig base rate for å gjenta kriminalitet.

Northpointe hevder at man heller skal se på «target population errors», som tar høyde for base-raten for kriminalitet blant afro-amerikanere og hvite. Positiv prediktiv verdi viser sannsynligheten for at en som ikke er klassifisert som lav-risiko for gjentagende kriminalitet, faktisk gjentar kriminalitet. Det komplementære til dette er at en som ikke er klassifisert som lav-risiko, ikke gjentar kriminalitet. Negativ prediktiv verdi viser sannsynligheten for at en som er klassifisert som lav-risiko for gjentagende kriminalitet faktisk ikke gjentar kriminalitet. Det komplementære til dette er at man faktisk gjentar kriminalitet når man har blitt klassifisert som lav risiko. Her mener både ProPublica og Northpointe at de er like nøyaktige.

Selv om det kan høres ut som at uenigheten mellom ProPublica og Northpointe kan ligge i deres statistiske tilnærming til å regne ut nøyaktigheten til modellen, kan dette løftes opp til et høyere nivå; uenigheten kan ligge i at ProPublica mener at algoritmen er diskriminerende, mens Northpointe mener at algoritmen har rett med utgangspunktet i det datagrunnlaget den har. Mellom linjene hevder altså Northpointe at det kanskje er datagrunnlaget som er skjevt. Dette viser tilbake til det vi har diskutert

tidligere i denne oppgaven om at det kan være vanskelig for algoritmer å gi rettferdige utslag hvis det allerede er eksisterende urettferdigheter i samfunnet.

Videre kan noe av kjernen i uenigheten mellom ProPublica og Northpointe også komme av deres juridiske eller moralske tilnærming til sensitiviteten av algoritmen. Er det slik at Northpointe er mer sensitiv fordi de ikke vil la alvorlige kriminelle gå fri? Og at ProPublica er mindre sensitive, fordi de ikke forgjeves vil dømme en uskyldig person. Dette reflekterer en gjennomgående diskusjon i rettsvesenet, hvor man ønsker å balansere individuelle rettigheter med samfunnets rettighet til å beskyttes mot kriminalitet. Dog, er et viktig juridisk prinsipp som er gjeldende i rettsvesenet at det er bedre å la en skyldig person gå fri hvis det ikke er bevist utover enhver rimelig tvil at personen er skyldig enn å gjøre en uskyldig person skyldig. Dette juridiske prinsippet burde også reflekteres i algoritmen når man nettopp bidrar til å avgjøre skyld eller uskyld i rettsvesenet.

4.3. Hvordan en tredjeparts algoritmerevisjon sett ut

Tidligere i oppgaven har vi sett på hvordan en rettferdig algoritme kan se ut, ved å både diskutere hva slags hensyn den må ta, hva slags virkninger den ideelt sett skal ha og hva den ikke skal ha. Med utgangspunkt i denne drøftingen skal vi skissere opp hvordan et system for en tredjeparts-algortimerevisjon kunne sett ut. En tredjeparts-algortimerevisjon er altså et uavhengig system, uavhengig av de som har laget algoritmen, som skal revidere hvordan en algoritme fungerer med tanke på å sikre rettferdig bruk og utfall av det i rettsvesenet.

Det første poenget som er viktig med en slik uavhengig tredjeparts-algortimerevisjon er at selv om den er uavhengig, krever det et samarbeid med blant annet de som har laget algoritmen. Uten dette samarbeidet kan for eksempel ikke denne tredjeparts- algortimerevisjonen få tilgang til de spørsmålene som inngår i algoritmen for å vurdere risikoen for å gjenta kriminaliteten på nytt. ProPublica har også belyst dette som et problem i sin artikkel, hvor de forteller at de må vurdere en algoritme uten å vite hvordan den egentlig vurderer. Vi vet noe om spørsmålene COMPAS modellen bygger sin vurdering på 137 spørsmål, slik som:

“Was one of your parents ever sent to jail or prison?” “How many of your friends/acquaintances are taking drugs illegally?” and “How often did you get in fights while at school?” The questionnaire also asks people to agree or disagree with statements such as “A hungry person has a right to steal” and “If people make me angry or lose my temper, I can be dangerous.”

Samtidig vet vi ikke hvordan disse ulike spørsmålene er vektet, og vi vet heller ikke det empiriske grunnlaget bak hvorfor disse spørsmålene er valgt. Som nevnt tidligere så er disse spørsmålene nettopp med på å belyse noe av skjevheten i datagrunnlaget, ved at det fanger opp risikofaktorer som er mer assosiert med afro-amerikanere enn hvite.

Debatten mellom ProPublica, en uavhengig, undersøkende avis, og Northpointe, selskapet som står bak algoritmen, er med på å belyse behovet for at det burde være felles retningslinjer for hvordan man skal vurdere om en algoritme er rettferdig eller ikke. Mens ProPublica for eksempel så mer på sensitiviteten og spesifisiteten til modellen, var Northpointe opptatt av at modellen hadde like nøyaktig treffsikkerhet for afro-amerikanere og hvite, og at den predikerende sannsynligheten burde ta utgangspunkt i base-raten for å gjenta kriminalitet. Det burde være tredjeparter som definerer felles kriterier for hva som er rettferdig eller ikke, og noe alle parter burde være enig om.

Et annet viktig poeng med tredjeparts- algortimerevisjonen er at det burde være transparente prosesser rundt både hvem som har vært med på å analysere dataen, hvordan de har analysert algoritmen, og hvordan de har rapportert funnene. Slik kan andre uavhengige tredjeparter også være med på å revidere denne revisjonen. Det kan hende det er visse begrensninger til hvor langt denne transparensen kan gå. For eksempel kan det være at selskaper for å beskytte sin algoritme ikke vil dele informasjon

om alle aspekter av den med alle sammen. Men da burde det komme fram i rapporten hvilke deler av algoritmen som ikke er med i rapporten, og refleksjoner rundt hvordan det er en metodologisk begrensing for revisjonen.

En annen viktig del av rapportene som disse tredjeparts- algortimerevisjonen skal gjøre er ikke bare å se på om en algoritme er rettferdig eller ikke, men hvis den finner funn som er med på å svekke rettferdigheten i rettsvesenet burde den komme med spesifikke mekanismer for hvordan man kan løse disse problemene. Det er her samarbeidet med de aktørene som har laget algoritmene igjen blir sentralt, fordi det kan være en kort vei fra en anbefaling om endring til at slike endringer faktisk blir implementert. Dette er direkte med på å øke den rettferdige bruken av algoritmer i rettsvesenet.

5. Konklusjon

Våre resultater var i tråd med Northpointe sine funn som viste like god treffsikkerhet for afro-amerikanere som for hvite. Våre funn kan derfor potensielt indikere at datagrunnlaget COMPAS har bygget på er det som er skjevt. Som diskutert tidligere i oppgaven er det vanskelig å lage en algoritme hvis det er allerede eksisterende skjevheter i samfunnet. Og det er også utfordringer knyttet til å prøve å kompensere for disse. Samtidig oppfordrer vi til bruk av tredjeparts algoritme revisjoner i fremtiden for å øke transparens rundt antagelsene og informasjonen algoritmen bygger på.

6. Referanser

Angwin, A., Larson, J., Mattu, S., & Kirchner, L. (2016, 23. mai). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1). https://doi.org/10.1177/2053951720983865

Dieterich, W., Mendoza, C., Brennan, T. (2016, 8. juli). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc. Research Department. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Han, B.-C. (2015). The Burnout Society. Stanford University Press.

Hughes, C. (2024). The End of Race Politics: Arguments for a Coloblind America. Penguin Random House.

Larson, J., Mattu, S., Kirchner, L., & Angwin, A. (2016, 23. mai). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

Ruha, B. (2019). Race After Technology. Polity Press.