Tidsskrift for velferdsforskning

Universitetsforlaget

VITENSKAPELIG PUBLIKASJON

Årgang 25, nr. 3-2022, s. 1–24
ISSN online: 2464-3076
DOI: https://doi.org/10.18261/tfv.25.3.3

Bias og kvantitativ analyse innen velferd

opphav til skjevheter og relasjon til utfallsrettferdighet

Bias in Quantitative Data Analysis for Welfare:

Origin of Biases and Relationship to Outcome Fairness

Andrea Marheim Storås
Doktogradsstipendiat ved Department of Holistic Systems, SimulaMet, Oslo, og Institutt for informasjonsteknologi, OsloMet
andrea@simula.no

Robindra Prabhu
Data scientist ved NAV IT Utvikling og Data, Arbeids- og velferdsdirektoratet, Oslo
robindra.prabhu@nav.no

Hugo Lewi Hammer
Professor i statistikk ved Institutt for informasjonsteknologi, OsloMet, og forsker II ved Department of Holistic Systems, SimulaMet, Oslo
hugo.hammer@oslomet.no

Inga Strümke
Forsker ved Institutt for teknisk kybernetikk, Norges teknisk-naturvitenskapelige universitet, Trondheim, og forsker II ved Department of Holistic Systems, SimulaMet, Oslo
inga@simula.no

Sammendrag

Ifølge Norges nasjonale strategi for kunstig intelligens er offentlig forvaltning og helse blant Norges satsingsområder for bruk av kunstig intelligens. Maskinlæring er en undergruppe av kunstig intelligens med potensial for å løse en rekke utfordringer, men som også gir opphav til utfordringer. En slik utfordring er bias – eller skjevhet. Et eksempel på skjevhet er at tilstedeværende ulikheter i samfunnet representeres i datagrunnlaget maskinlæringsmodeller utvikles på. De resulterende modellene står dermed i fare for å adoptere og videreføre disse ulikhetene. En utfordring er at skjevhet har ulike definisjoner innen ulike fagområder, og kan ha mange ulike opphav. Vi bidrar til å løse denne utfordringen ved å gi en oversikt over ulike typer skjevhet og deres opphav med illustrasjoner fra et velferdsperspektiv, og vi avklarer forskjellen til det nærliggende konseptet rettferdighet. Vi demonstrerer utfordringer relatert til databaserte modellers oppførsel ved å benytte maskinlæring til å predikere fremtidig ressursbehov i helsevesenet, spesifikt antall legebesøk i kommuner. Vi demonstrerer ulike typer skjevheter, diskuterer mulige løsninger og bruker metoder fra forklarbar kunstig intelligens for å analysere opphavet til skjevheter i forklaringsvariablene. Det finnes ingen universell løsning for å håndtere alle typer skjevheter, men skjevhet må tas høyde for i alle deler av en kvantitativ analyse.

Nøkkelord
XGBoost, forklarbar kunstig intelligens, velferdsforskning, maskinlæring, skjevhetsbegreper

Abstract

According to Norway's national strategy for artificial intelligence, public administration and health are among Norway's focus areas within artificial intelligence. Machine learning, a branch of artificial intelligence, has strong potential for solving problems, but simultaneously creates challenges. One such challenge is bias. If underlying inequalities in society are represented in the data used to develop machine learning models, the resulting models are at risk of adopting and perpetuating these inequalities. Furthermore, bias has different definitions within different scientific fields and can result from a variety of different causes.

We contribute to solving this by providing an overview of different types of biases and their origins, accompanied by illustrations from a welfare context. We highlight how bias differs from the related concept of fairness. We demonstrate challenges related to the behaviour of data-based models by using machine learning to predict resource demands in healthcare. We demonstrate different types of biases, discuss solutions and apply methods from explainable artificial intelligence to analyse the origin of biases in the model features. In conclusion, there is no universal solution for handling all types of biases, and bias must be taken into account in every part of a quantitative data analysis.

Keywords

XGBoost, explainable artificial intelligence, welfare research, machine learning, bias terms

1 Introduksjon

I denne artikkelen analyserer vi fenomenet som er betegnet av det engelske begrepet *bias* innen datainnsamling og kvantitativ analyse, som vi velger å oversette til *skjevhet*. Skjevhet kan oppstå som følge av mange ulike årsaker og kommer i mange varianter. Det finnes ingen definisjon i litteraturen som dekker alle formene for skjevhet, men overordnet kan det beskrives som *resultater eller slutninger som systematisk avviker fra de virkelige forholdene som utforskes i en studie* (Grønmo, 2020). Innen samfunnsforskning finnes det flere ulike definisjoner av skjevhet (Aronson et al., 2021). Hammersley og Gomm (1997) foreslår å definere det som en systematisk feil som burde vært oppdaget og minimert[1]. I Centre for Data Ethics and Innovation (2020) defineres begrepet som et utfall som ikke bare er skjevt, men forskjøvet slik at utfallet er urettferdig.

En norsk oversikt over ulike typer skjevhet er utarbeidet av Staff (2015). Vår artikkel skiller seg fra denne ved at vi beskriver skjevheter som er spesielt relevante i utviklingen og anvendelsen av maskinlæringsmodeller, en undergruppe av kunstig intelligens der dataprogrammer lærer å oppnå et mål basert på data. For å illustrere hvordan skjevhetene kan komme til uttrykk, inkluderer vi hypotetiske eksempler fra Arbeids- og velferdsetaten (NAV)[2]. Skjevhet er en velkjent utfordring innen innsamling og analyse av data. Maskinlæringsmodeller skiller seg imidlertid fra tradisjonelle analysemetoder ved at de oftere brukes til beslutningstaking, og sjeldnere til å finne årsakssammenhenger mellom variabler. Dette fører til at skjevhet har fått fornyet aktualitet i offentlig diskurs.

I senere år har media rapportert om såkalt *algorithmic bias* (Centre for Data Ethics and Innovation, 2020) i databaserte modeller anvendt på en rekke ulike områder. Debatten nådde norsk offentlighet sommeren 2020, da avgangskarakterer for elever på IB-linjen ble satt av en databasert modell (Datatilsynet, 2020). Både Nasjonal strategi for kunstig intelligens (Kommunal- og moderniseringsdepartementet, 2020) og EU-kommisjonens ekspertgruppe innen kunstig intelligens (Independent High-Level Expert Group on Artificial Intelligence, 2019) peker på skjevhet som en utfordring som må unngås ved bruk av kunstig

1. Oversatt av artikkelforfatterne.
2. Illustrasjonene er valgt for å billedliggjøre ulike former for skjevheter som kan oppstå i utviklingsprosessen, og viser ikke til systemer som er produksjonssatt eller planlagt produksjonssatt av NAV.

intelligens og databaserte modeller. I et nytt lovforslag fra EU (European Commission, 2021) vil trolig mange anvendelser av databaserte modeller i offentlig forvaltning og på velferdsområdet underlegges strengere krav hva gjelder skjevheter i datagrunnlag og modellutvikling.

Uten verktøy for kontroll og styring kan data og databaserte modeller videreføre, forsterke og befeste skjev og uønsket praksis – i verste fall tilslørt som nøytrale «faktabaserte» systemer. Riktige verktøy kan bidra til å avdekke eksisterende skjevheter og utvikle tiltak. Skjevhetsutfordringen er mangefasettert og har ofte både sosiale, normative og tekniske dimensjoner. Ulike fagdisipliner belyser problemstillingen på forskjellige måter. Samtidig er begrepene sammenvevd: En statistisk skjevhet kommer sjelden uten en normativ slagside, og en sosial skjevhet vil også vises i statistiske mål. For å adressere problemet må begge sider forstås.

I likhet med skjevhet har diskusjonen om urettferdige maskinlæringsmodeller fått stor oppmerksomhet de siste årene (Bolukbasi et al., 2016; Bellamy et al., 2018; Gianfrancesco et al., 2018), noe som har gitt opphav til et nytt og tverrfaglig forskningsfelt. Ulike mål på rettferdighet, på engelsk *fairness* (Verma & Rubin, 2018), er foreslått brukt i en maskinlæringskontekst. Flere av disse kan knyttes til ulike moralteoretiske ideer om (utfalls)rettferdighet (Barocas et al., 2019; Verma & Rubin, 2018) og uttrykkes som betingede uavhengigheter mellom tre variabler: (1) en *sensitiv* egenskap S man vil beskytte, (2) den forklarte variabelen Y og (3) tapskriteriet U.

Flere oversiktsartikler om skjevhet og maskinlæring har blitt publisert, se for eksempel Srinivasan og Chander (2021); Mehrabi et al. (2021); Suresh og Guttag (2021) og Gianfrancesco et al. (2018). Noen inkluderer også rettferdighetsaspekter (Mehrabi et al., 2021; Gianfrancesco et al., 2018), men de tar ikke for seg relasjonen til skjevhet. Vår artikkel skiller seg fra tidligere artikler ved at vi diskuterer relasjonen mellom skjevhet og rettferdighet og ser på hvordan dette kan være relevant i en velferdskontekst.

Artikkelen er organisert som følger: I seksjon 2 beskriver vi opphavet til ulike skjevheter og presenterer ulike definisjoner fra litteraturen. Seksjon 3 illustrerer ulike skjevheter gjennom et eksempel som predikerer fremtidig ressursbehov i helsevesenet. I seksjon 4 diskuteres forholdet mellom rettferdighet og skjevhet. Vi avslutter med en diskusjon av funnene og trekker konklusjoner i seksjon 5.

2 Skjevhet

I dette avsnittet diskuterer vi ulike skjevheter og hvor de kan oppstå i kvantitative analyser. En samlet oversikt vises i figur 1. Vedlegg A.1 diskuterer norske oversettelser av begrepet *bias* og de ulike skjevhetene.

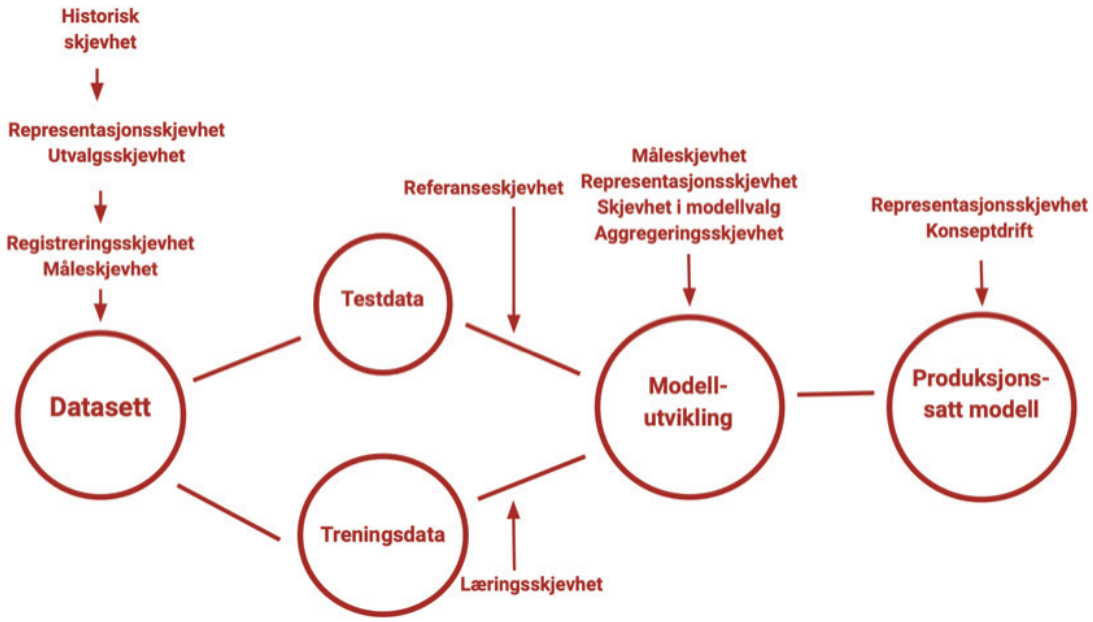
2.1 Skjevhet i datagrunnlaget

Datagrunnlaget som benyttes til utvikling av maskinlæringsmodeller, kan gi opphav til flere former for skjevhet, og vi trekker frem et relevant utvalg.

Historisk skjevhet

Historiske data beskriver situasjonen da dataene ble samlet inn. Siden data kan inneholde grupper som av ulike årsaker har hatt mindre tilgang til eller gjort mindre krav på ressurser, er det stor fare for at modeller utviklet på slike data, propagerer uønsket praksis (Bolukbasi et al., 2016).

Følgende er en tenkt situasjon hos NAV: NAV utvikler en løsning som anbefaler ledige stillinger til registrerte arbeidssøkere dels basert på historiske arbeidsforhold. Hvis løsningen anbefaler fortrinnsvis kvinner til historisk kvinnedominerte yrker, og likeledes for menn, utviser den historisk skjevhet.



Figur 1: Oversikt over ulike kilder til skjevhet som kan oppstå i løpet av en kvantitativ analyse. Figuren er inspirert av Suresh og Guttag (2021).

Registreringsskjevhet

Registrerte data gir ideelt sett en sannferdig fremstilling av en tilstand på registreringstidspunktet. Den sosiale konteksten for registreringen kan likevel påvirke hvilken informasjon som registreres. I motsetning til historisk skjevhet kan registreringsskjevhet oppstå i reell tid og ikke som følge av historiske endringer i samfunnet.

En tenkt situasjon hos NAV der registreringsskjevhet kan oppstå, er hvis noen veiledere benytter fritekstfelt for å supplere med tilleggsinformasjon om sykefravær, mens andre ikke gjør det. Da oppstår det en forskjell mellom de to registreringsformene.

Måleskjevhet

Vi oversetter det engelske begrepet *measurement bias* (Page & Henderson, 2008) til måleskjevhet. Dette beskriver en hvilken som helst systematisk eller tilfeldig målefeil som skjer under innsamling av data, hvor det målte ikke er direkte observerbart. Dataene blir dermed en tilnærming eller overforenkling. Måleskjevhet er relevant for utvikling av maskinlæringsmodeller fordi variablene som inkluderes i modellen, kan være *proxyer* (Suresh & Guttag, 2021), altså at de ikke beskriver direkte årsaksforhold, men inneholder indirekte informasjon om det (Dhrymes, 2017).

Vi tenker oss at NAV vil bistå veiledere i vurderingen av oppfølgingsbehov for arbeidssøkere med «hull i CV-en» ved å gi en indikasjon på hvor langt de står fra arbeidsmarkedet. De utvikler et mål på «avstand til arbeidsmarkedet» som blant annet benytter tidligere pen-

sjonsgivende inntekt. «Avstand til arbeidsmarkedet» er imidlertid ikke trivielt oversatt til en målbar metrikk og avhenger av mange faktorer. «Tidligere pensjonsgivende inntekt» som en indikator for arbeidserfaring omfavner ikke all arbeidserfaring og kan treffe arbeidssøkere med arbeidserfaring fra utlandet skjevt.

Representasjonsskjevhet

Det engelske begrepet *representation bias* (Suresh & Guttag, 2021), som vi oversetter til representasjonsskjevhet, kan oppstå som følge av ulike mekanismer og på ulike steder i datainnsamling og kvantitativ analyse, som er vist i figur 1. Representasjonsskjevhet gjør seg gjeldende som en ikke-representativ populasjon i dataene modellen trenes på, sammenliknet med populasjonen den benyttes på. Representasjonsskjevhet kan skyldes at utvelgelsen ikke er tilfeldig, eller at utvalgspopulasjonen ikke samsvarer med populasjonen modellen brukes på (Suresh & Guttag, 2021).

Følgende er en tenkt situasjon hos NAV: Etter å ha produksjonssatt en maskinlæringsmodell, høster NAV gode erfaringer med et arbeidsmarkedstiltak i to fylker. Resultatene benyttes i en nasjonal anbefalingsløsning for arbeidssøkere med tilretteleggingsbehov, uten at tiltaket vurderes mot lokale arbeidsmarkeder i andre fylker.

Utvalgsskjevhet

Utvalgsskjevhet er et spesialtilfelle av representasjonsskjevhet og oppstår fordi utvelgelsen av observasjoner ikke er tilfeldig. For databaserte modeller betyr det at dataene de utvikles på, ikke er representative for dataene de skal brukes på (Heckman, 1979). Dette er aktuelt for all modellering av data, ikke bare maskinlæring.

Vi ser for oss at for å kartlegge hvilke tema som opptar brukere, gjennomfører NAV en analyse av «skriv-til-oss»-henvendelser, der henvendelsene grupperes etter innhold. Dermed forbedres innholdet for å imøtekomme behovene til brukere som aktivt benytter seg av «skriv-til-oss»-tjenesten, men ikke til brukere som av ulike grunner ikke benytter tjenesten.

2.2 Skjevhet innen dataanalyse og modellutvikling

Innen statistikk defineres skjevhet som avvik mellom forventningsverdien til en estimator og verdien til parameteren som estimeres (Cox, 2006) og omtales som *forventningsskjevhet*. Forventningsskjevhet i prediksjoner oppstår hvis modellen ikke er egnet til å beskrive virkeligheten basert på data (James et al., 2013), for eksempel hvis en lineær regresjonsmodell benyttes på data med uttalte ikke-linearitet er.

Maskinlæring favner optimaliseringsalgoritmer som lærer å oppnå mål basert på data. All skjevhet som er relevant for statistisk modellering, er derfor også relevant for maskinlæring. Mitchell (1980) definerer skjevhet innen maskinlæring som «et hvilket som helst grunnlag for at en generalisering velges over en annen, annet enn en perfekt gjengivelse av de observerte dataene»[3].

Vi bruker følgende notasjon for å beskrive data for modellutvikling: Et datasett består av ulike forklaringsvariabler, X, og en eller flere forklarte variabler, Y. Dataene som benyttes for å utvikle modellen, kalles *treningsdata*, mens modellen evalueres på *testdata*, som ikke var tilgjengelig for modellen ved trening. En maskinlæringsmodell trenes ved at den minime-

3. Oversatt av artikkelforfatterne.

rer et tapskriterium, som representerer avviket mellom modellens prediksjon og den sanne verdien i treningsdatasettet[4], og bestemmes av modellutvikleren.

Skjevhet i modellvalg

Når maskinlæringsmodeller utvikles, er det vanlig å prøve flere modeller og velge den beste basert på forhåndsbestemte kriterier. Valg av kriterier er subjektivt og kan derfor være en kilde til skjevhet (Choi et al., 2019). Så lenge det ikke kan bestemmes hvorvidt en maskinlæringsmodell er den best mulige modelleringen av treningsdataene, vil alle maskinlæringsmodeller ifølge Mitchell (1980) ha skjevhet.

Vi ser for oss at NAV utvikler en modell for å flagge hvilke saker i en automatiseringssløyfe som bør tas ut til manuell behandling. To ulike modellarkitekturer (A og B) prøves ut, og modell A velges på bakgrunn av dens enkelthet og tolkbarhet. Modell A har imidlertid en høyere falsk positiv-rate for sykemeldte med psykiske sykdomsbilder, slik at denne gruppen oftere feilaktig tas ut til kontroll enn ved bruk av modell B.

Læringsskjevhet

Nært beslektet til skjevhet i modellvalg er skjevhet i valg av modellens tapsfunksjon. Vi oversetter slik *learning bias* (Suresh & Guttag, 2021) til læringsskjevhet. Når en modell optimaliserer én tapsfunksjon, skjer det på bekostning av andre tapskriterier (Kleinberg et al., 2016). Dette oppstår når optimalisering skjer med hensyn til et helt datasett og går på bekostning av underrepresenterte grupper i dataene (Hardt et al., 2016).

Vi forestiller oss at NAV utvikler en modell for å predikere lengden på sykefravær og trener modellen for å oppnå best mulig treffsikkerhet. På bekostning av dette presterer modellen litt dårligere på menn enn kvinner.

Referanseskjevhet

Ofte benyttes ferdige referansedata for å evaluere modeller (Gijsbers et al., 2019). Dette kan føre til såkalt referanseskjevhet, på engelsk *evaluation bias* (Suresh & Guttag, 2021), ved at modellene som gjør det best på referansedataene, foretrekkes uten at disse generaliserer best til dataene modellen brukes på etter produksjonssetting. Et kjent eksempel er ansiktsgjenkjenningsmodeller som evalueres på referansedata med få tilfeller av mørkhudete kvinner, men som senere benyttes for blant annet denne gruppen, og viser seg å ikke fungere (Buolamwini & Gebru, 2018). Slik skjevhet er mindre aktuelt hos NAV, da ferdige referansedata sjelden vil benyttes til evaluering.

Aggregeringsskjevhet

Aggregeringsskjevhet, på engelsk *aggregation bias*, brukes av Feige og Watts (1972) for å beskrive at modellen gjør feilaktige antagelser om individer eller minoritetspopulasjoner i datagrunnlaget, gjennom ekstrapolering fra en majoritetspopulasjon. Slik skjevhet kan gjøre at modellen underpresterer for alle undergrupper, eller at den fungerer best på undergruppen med flest observasjoner.

4. Vi begrenser oss til veiledet læring, hvor treningsdatasettet inneholder den sanne verdien.

Vi tenker oss at NAV utvikler en modell for å predikere lengden av sykefravær. Modellen presterer dårlig på brukere som er sykemeldt grunnet svangerskapsrelaterte plager. Nærmere undersøkelser viser at underdiagnoser i denne gruppen gir ulike fraværslengder. Siden modellen ser på alle brukere med svangerskapsrelaterte plager samlet, fanges ikke dette opp.

Konseptdrift

Konseptdrift, på engelsk *concept drift,* ble introdusert av Schlimmer og Granger (1986) for å beskrive skjevhet som følge av endringer over tid, hvor relasjonen mellom forklaringsvariablene og den forklarte variabelen endres. Dersom en databasert modell ikke tar høyde for dette, vil ytelsen gradvis forverres når endringene blir større. Konseptdrift kan være vanskelig å detektere, da eventuelle observerte endringer også kan skyldes legitime effekter som støy eller uteliggere i dataene (Žliobaitė et al., 2016).

En tenkt situasjon er at NAV produksjonssetter en modell for å predikere sykefraværsvarighet. Modellen benytter blant annet diagnosekoder fra leger. Praksis med utmåling av visse diagnoser i legemeldte sykefravær endres over tid, uten at modellen endres.

3 Illustrasjon gjennom legekonsultasjoner

Vi illustrerer skjevhet – og senere relasjonen til utfallsrettferdighet – gjennom prediksjon av fremtidig ressursbehov i helsevesenet. Med *prediksjon* menes en prognose gitt av en modell basert på data, uavhengig av om dataene representerer fortid, nåtid eller fremtid. Vi bruker data fra NAV (Goth et al., 2014) og illustrerer skjevheter som kan oppstå, samt hvordan de kan korrigeres. Mye forskning er viet prediksjon av fremtidige ressursbehov i helsevesenet, se for eksempel Ordu et al. (2021).

God prediksjon av behovet for fremtidige legekonsultasjoner i ulike geografiske områder kan være et nyttig verktøy for å planlegge og allokere helseressurser. Vi utvikler maskinlæringsmodeller som predikerer antall legebesøk i løpet av en uke, basert på data fra tre foregående uker. Fordi personer med alder ≥ 70 år er en sårbar gruppe med behov for god legetilgang, predikerer vi antall ukentlige konsultasjoner for denne gruppen, basert på data fra alle aldersgrupper. Dette kan naturligvis utvides til andre pasientgrupper. Vi viser prediksjoner for Oslo og Kragerø, som representerer henholdsvis en stor og en liten kommune. Spesielt forventer vi at en liten kommune vil ha nytte av data fra flere kommuner for å øke den totale datamengden.

Maskinlæringsmodellene vi benytter i eksemplene, er mindre komplekse enn dyplæringsmodeller, men de belyste problemstillingene er modellagnostiske, da mer komplekse modeller basert på mer data kan gi opphav til liknende situasjoner. I en velferdskontekst kan mer data også øke risikoen for introduksjon av skjevheter fordi observasjoner samles inn med flere variabler og over lengre tid, og kan i større grad fange opp samfunnsendringer. Læringsalgoritmene vi bruker, er aktuelle for kvantitativ analyse av strukturerte data (Géron, 2019) – en vanlig datatype i en velferdskontekst. Detaljer om dataene og valg av maskinlæringsalgoritmer er vedlagt (vedlegg A.2).

3.1 Modellevaluering

Prediksjon av antall legekonsultasjoner er et regresjonsproblem, og vi bruker kvadratroten av kvadrert avvik mellom modellprediksjon og observert verdi – root mean squared error (RMSE) – som mål på modellens treffsikkerhet. Den rapporterte treffsikkerheten gjenspeiler ikke nødvendigvis treffsikkerheten for hendelser som ennå ikke har oppstått.

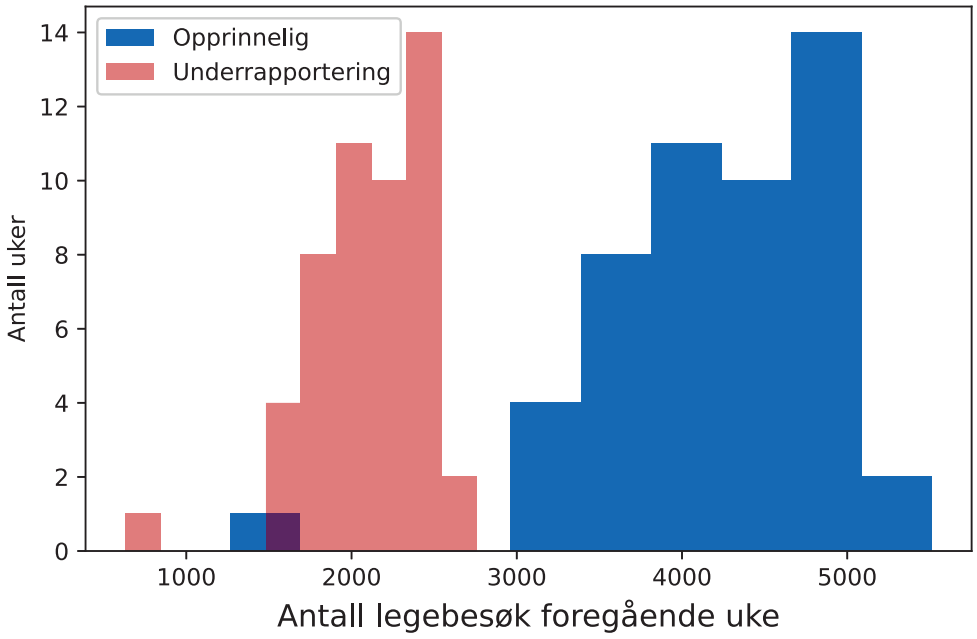
Vi ønsker også å vite hvilke variabler som er viktige for modellen, og benytter forklaringsmetodene Shapley additive explanations (SHAP) (Lundberg et al., 2019) og Shapley additive global importance (SAGE) (Covert et al., 2021), se detaljer i vedlegg A.3. SHAP-verdier indikerer i hvilken retning de ulike modellvariablene trekker prediksjonen, mens SAGE-verdier indikerer hvor mye hver variabel bidrar til modellens totale tap. Begge metodene gir en rangering av variablenes viktighet i prediksjonsproblemet.

Modellene evalueres på testdata fra samme distribusjon som treningsdataene. Det er mye som kan føre til at dette ikke er tilfellet for produksjonssatte modeller. Vi gir eksempler på skjevhet som følge av underrapportering (seksjon 3.2) og endringer over tid (vedlegg A.7). Et utvalg metoder som kan benyttes for å korrigere slike former for skjevhet, beskrives i vedlegg A.4.

3.2 Skjevhet grunnet underrapportering

For å studere effekten av skjevhet i forklaringsvariabler innfører vi forsinkelser i rapportering av legekonsultasjoner og undersøker oppførselen til modellene (se detaljer i vedlegg A.5).

Antall konsultasjoner for foregående uke reduseres til 0,5 av opprinnelig antall i testdataene. Figur 2 viser opprinnelig og forskjøvet fordeling av konsultasjoner uken før prediksjonstidspunktet for Oslo kommune (kommunenummer 301). Dette utgjør en utvalgsskjevhet, siden det gir systematisk avvik i noen forklaringsvariabler. Det kan også anses som konseptdrift fordi relasjonen mellom forklaringsvariablene og den forklarte variabelen endres. For å studere effekten av skjevheten lar vi modellene predikere på testdatasett både med og uten skjevhet.



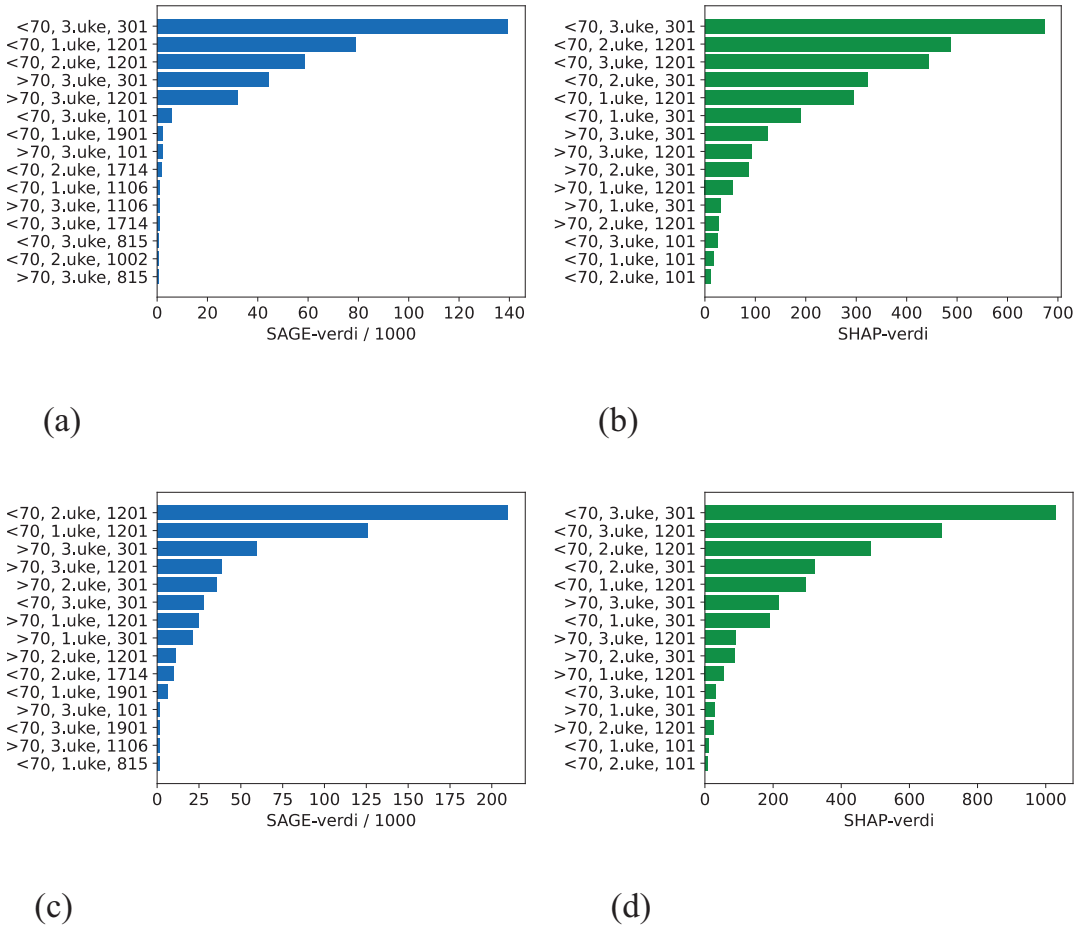
Figur 2: Fordelingen av ukentlige legekonsultasjoner for personer ≥ 70 år i Oslo kommune uken før prediksjonstidspunktet. Konsultasjonstall for 21 ulike uker er inkludert. Blå søyler representerer opprinnelig antall legekonsultasjoner, og røde representerer antall konsultasjoner etter reduksjon til halvparten av opprinnelig antall.

3.2.1 Oslo kommune

For Oslo kommune øker RMSE fra 750 til 1446 når skjevhet innføres i testdatasettet. Videre beregnes SAGE- og SHAP-verdier. Før innføring av skjevhet er antall legekonsultasjoner for personer < 70 år i Oslo den foregående uken den viktigste variabelen for modellen, se figu-

rene 3a og 3b. Videre er antall konsultasjoner i Bergen (kommunenummer 1201) i ulike uker viktige. Den mest sannsynlige grunnen til at gruppen < 70 år er viktig for modellen, er at det er relativt flere personer i denne gruppen. Det er også mulig at noen sykdommer rammer de yngre i samfunnet først, før de sprer seg til personer ≥ 70 år, men dette er vår spekulasjon. At Oslo og Bergen er byene med flest innbyggere, kan forklare hvorfor tall fra disse byene vektlegges av modellen. Bergen har sannsynligvis flere likhetstrekk med Oslo fordi begge er store byer, slik at trender i Bergen kan gjenspeile trender i Oslo.

Etter innføring av skjevhet endrer rekkefølgen til de tre viktigste forklaringsvariablene seg for både SAGE- og SHAP-verdiene, se figurene 3c og 3d. Ifølge SAGE-verdiene er antall legekonsultasjoner for personer < 70 år i Bergen viktigst for modellen, mens konsultasjonstallene den foregående uken for personer ≥ 70 år i Oslo rykker ned til tredjeplass. For SHAP-verdiene har forklaringsvariablene på andre- og tredjeplass byttet plass.



Figur 3: (a) SAGE-verdier og (b) SHAP-verdier for prediksjoner på personer med alder ≥ 70 år i Oslo kommune i 2010. (c) Tilsvarende SAGE-verdier og (d) SHAP-verdier ved kunstig lave besøkstall for foregående uke.

3.2.2 Kragerø kommune

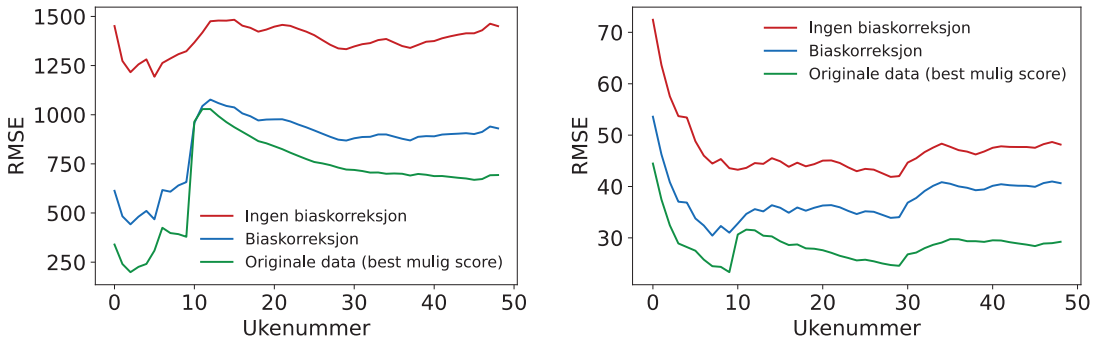
Etter innføring av samme skjevhet som i seksjon 3.2.1 øker RMSE fra 30 til 48. Kragerø har langt færre innbyggere og legekonsultasjoner enn Oslo, noe som gjenspeiles i lavere predikerte verdier og RMSE. Ifølge SAGE- og SHAP-verdiene er konsultasjonstall for personer bosatt i Oslo (301) og Bergen (1201) de viktigste forklaringsvariablene for modellen

både før og etter innføring av skjevhet, se vedlagte figurer S2a til S2d. At modellen anser antall legekonsultasjoner fra Oslo og Bergen som viktige for å predikere konsultasjonstall i Kragerø, skyldes sannsynligvis at kommunene er store og bidrar med mye data. Som for prediksjonene i seksjon 3.2.1 endres SAGE- og SHAP-verdiene etter innføring av skjevhet. Rekkefølgen til de tre viktigste forklaringsvariablene er endret ifølge SAGE-verdiene, mens den er uendret ifølge SHAP-verdiene. For SAGE-verdiene rykker konsultasjonstall for Bergen den foregående uken ned fra andre- til tredjeplass. Reduksjonen av konsultasjonstallene for den foregående uken kan forklare hvorfor dette blir en mindre viktig forklaringsvariabel for modellen, ifølge SAGE- verdiene, som baserer seg på modellens totale tap.

3.3 Korreksjon av underrapportering

For å oppdage skjevheter i forklaringsvariablene studerer vi forskjellen i fordelingene for trenings- og testsett for hver variabel. Skjevheter kan korrigeres ved å justere forklaringsvariablene i testsettet som avviker systematisk fra treningssettet. Modellens prediksjoner på korrigerte data sammenliknes med tilsvarende prediksjoner gjort på de skjeve dataene uten korreksjon, og vi lar til sammenlikning modellen predikere på de originale dataene fra 2010. Se vedlegg A.6 for detaljer.

Resultatene for henholdsvis Oslo og Kragerø er vist i figurene 4a og 4b. I begge tilfeller er feilene lavest for data uten skjevhet (grønn kurve), mens de er størst for dataene med skjevhet (rød kurve). Når de skjeve dataene korrigeres og modellen brukes på disse (blå kurve), reduseres feilene. Effekten er tydeligst for Oslo. RMSE øker brått ved uke 11 både for originale data uten skjevhet og de skjeve dataene etter korreksjon. Trenden fanges derimot ikke opp for de skjeve dataene uten korreksjon. Nærmere undersøkelser viser at antall konsultasjoner falt betraktelig for denne uken, som sannsynligvis sammenfalt med påsken. RMSE-verdiene for prediksjoner på de skjeve datasettene etter korreksjon synker fra 1446 til 937 for Oslo og fra 48 til 41 for Kragerø. Korreksjonsmetoden er ikke perfekt, og de blå kurvene ligger høyere enn de grønne kurvene. En årsak er at metoden tar utgangspunkt i gjennomsnittsverdiene fra historiske data for å korrigere data i sanntid. Dette fører til konseptdrift, da underliggende trender i dataene kan ha endret seg over tid. Korreksjon av konseptdrift beskrives i vedlegg A.7.



Figur 4: Prediksjonsfeil for konsultasjonstall for personer med alder ≥ 70 i Oslo (a) og Kragerø (b). Grønne kurver tilsvarer prediksjoner gjort på data uten skjevhet, blå kurver tilsvarer korrigerte data, og røde kurver tilsvarer data med skjevhet uten korreksjoner. RMSE: Root Mean Squared Error

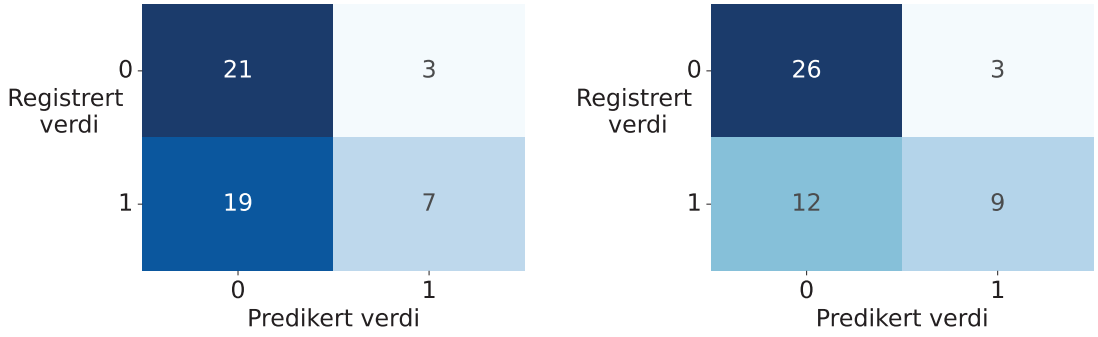
4 Rettferdighet

Dette avsnittet tar for seg skillet mellom utfallsrettferdighet og skjevhet og viser konkrete eksempler som illustrerer noen forskjeller.

I offentlig diskurs er en urettferdig modell gjerne ensbetydende med en skjev modell, men selv om rettferdighet og skjevhet er tematisk beslektet, representerer de ulike statistiske begreper. Vi presiserer her hvordan et populært ønskemål for rettferdighet skiller seg fra skjevhet slik sistnevnte er definert i denne artikkelen: En maskinlæringsmodell sies å være rettferdig overfor en egenskap (S) når den ikke favoriserer eller forhåndsdømmer individer med den gitte egenskapen (Mehrabi et al., 2021). Hvis en modell for eksempel predikerer ulikt for to testpersoner som er identiske utover at de tilhører ulike kjønn, er modellen i konflikt med ovennevnte rettferdighetskriterium med hensyn til egenskapen «kjønn». Innen statistikk omtales dette som betinget uavhengighet, se for eksempel Mary et al. (2019), og skiller seg fra skjevhet slik det her er omtalt: Rettferdighet er et formelt krav som modellen enten oppfyller eller ikke, mens skjevhet er en systematisk forskyvning i datagrunnlag, modell eller anvendelse.

Skjevhet og utfallsrettferdighet kan likevel vekselvirke og være relevante samtidig. For å illustrere en mulig relasjon mellom skjevhet og utfallsrettferdighet presenterer vi her modellprediksjonene i seksjon 3.2 som binære klasser i forvirringsmatriser, en viktig evalueringsmetrikk innen maskinlæring. Matrisene består av fire felt, hvor antall riktige prediksjoner («sanne positive» og «sanne negative») befinner seg i henholdsvis nedre høyre og øvre venstre hjørner, mens feilaktige prediksjoner («falske positive» og «falske negative») befinner seg i øvre høyre og nedre venstre hjørner.

Legebesøk i kommende uke konverteres til en binær variabel som tar verdien 0 eller 1 dersom antall konsultasjoner er henholdsvis lavere enn, større enn eller lik den foregående uken. Tilsvarende gjøres for modellprediksjonene av antall legebesøk for kommende uke. Ulike former for skjevhet innføres, og effekten på forvirringsmatrisene observeres. Vi fokuserer på falske negative, siden dette kan føre til at ressursbehovet underestimeres, og at innbyggerne i en kommune ikke får et helsetilbud som dekker deres behov. Kriteriet for utfallsrettferdighet er således at falsk negativ rate (FNR) er lik for alle kommuner. Forvirringsmatrisene for Oslo og Kragerø kommune uten skjevhet er vist i henholdsvis figur 5a og figur 5b. FNR er 19 / (19 + 7) = 0,73 for Oslo og 12 / (9 + 12) = 0,57 for Kragerø. Ratio mellom FNR i Oslo og Kragerø blir 1,28. Ratioen avviker fra 1 og illustrerer at modellen har en høyere relativ feilrate for Oslo, som kan oppfattes som urettferdig, selv uten innføring av skjevhet i dataene.



Figur 5: Forvirringsmatriser uten skjevhet for (a) Oslo og (b) Kragerø kommune.

Følgende fire former for underrapportering studeres: underrapportering for alle aldre, som beskrevet i seksjon 3.2, og underrapportering som rammer henholdsvis kun Oslo kommune, personer < 70 år og personer < 30 år. I alle scenariene beregnes FNR for prediksjoner av legebesøk i Oslo og Kragerø og FNR-ratio. Resultatene er oppsummert i tabell 1 og viser at alle formene for underrapportering gir endret FNR-ratio. Underrapportering i alle aldersgrupper treffer ulikt i de to kommunene og fører til høyere, men også likere FNR. SAGE- (figurene 3c og S2c) og SHAP-verdiene (figurene 3d og S2d) for Oslo- og Kragerø-modellene viser at forklaringsvariablene vektes ulikt, hvilket forklarer hvorfor underrapporteringen gir ulike utfall for de to kommunene.

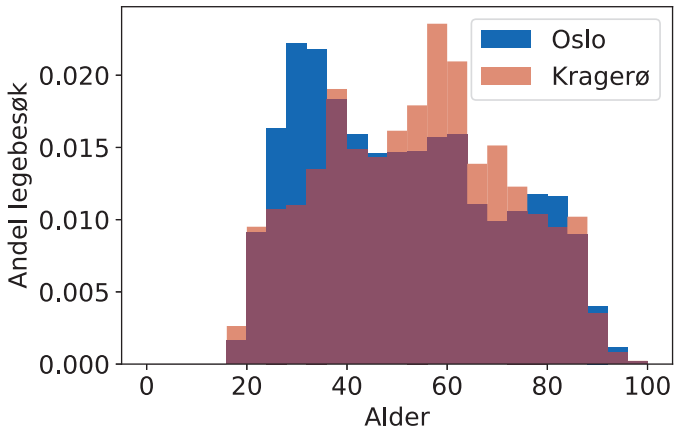
Tabell 1: Oversikt over FNR ved innføring av ulike former for underrapportering. Ratio angir forholdet mellom FNR for Oslo og Kragerø.

Type underrapportering	FNR Oslo	FNR Kragerø	Ratio
Ingen underrapportering	0,73	0,57	1,28
Alle aldersgrupper	0,92	0,86	1,07
Kun Oslo	0,96	0,95	1,01
Kun < 70 år	0,85	0,67	1,27
Kun < 30 år	0,81	0,57	1,42

FNR-ratio mellom Oslo og Kragerø etter innføring av skjevhet for personer < 70 år endres ikke nevneverdig, men den øker når skjevheten kun innføres for personer < 30 år. Fordelingen av legebesøk i de to kommunene for ulike aldersgrupper i 2006 og 2007 vises i figur 6. Andelen legebesøk for de < 70 år er relativt lik for de to kommunene, mens andelen for de < 30 år er større i Oslo enn i Kragerø. Dette viser at underliggende forskjeller i fordelingene kan føre til at skjevheter treffer ulikt for ulike grupper. At svært få legebesøk i Kragerø er fra personer < 30 år, kan forklare hvorfor underrapportering i denne aldersgruppen ikke påvirker FNR.

Eksemplene viser at skjevhet kan føre til utfall som *kan* oppfattes som urettferdige, men urettferdige utfall kan også følge av ulikheter i grunnfordelinger mellom ulike grupper som ikke nødvendigvis må tilskrives skjevheter. Skjevheter og utfallsrettferdighet vekselvirker gjennom underliggende fordelinger: Dersom aldersfordelingen er lik for Oslo og Kragerø, vil innføring av skjevhet i en aldersgruppe treffe relativt likt i de to kommunene. Ulike aldersfordelinger kan derimot føre til at skjevhet treffer ulikt.

I likhet med skjevhet er rettferdighetsbegrepet ofte normativt og kontekstuelt betinget, og matematiske definisjoner vil sjelden være dekkende. Relasjonen mellom skjevhet og rettferdighet beskrives ytterligere i vedlegg A.8.



Figur 6: Fordeling av legekonsultasjoner i ulike aldersgrupper i 2006 og 2007 for Oslo og Kragerø. Blå søyler representerer konsultasjoner i Oslo, og røde søyler representerer Kragerø.

5 Diskusjon og konklusjoner

I denne artikkelen kartlegger vi skjevhetsbegrepet innenfor rammene av utvikling av maskinlæringsmodeller til bruk i en velferdskontekst. Vi viser at det finnes mange former for skjevheter, med ulike opphav og virkninger, med illustrasjoner av hypotetiske situasjoner hos NAV. Vi viser også at forklaringsmetoder basert på Shapley-verdier, viser hvilke variabler modellen oppfatter som viktige, og som dermed er særlig sårbare for skjevheter.

Vi har illustrert virkninger og et lite utvalg korreksjonsmetoder på avidentifiserte registerdata, dog har vi presentert relativt enkle eksempler og diskutert de ulike formene for skjevhet isolert. Vi anser dette som den største begrensningen i vår analyse, da vi i virkeligheten må forvente at flere typer skjevheter, med sine ulike kilder, spiller sammen og er vanskelig å avdekke, skille fra hverandre og korrigere for. Korreksjon av én type skjevhet kan svekke modellens prestasjoner eller gjøre andre skjevheter større. Det bør gjøres veloverveide avveininger rundt hvilke korreksjoner som eventuelt skal benyttes, og hvilke konsekvenser det får. Som følge av plassmangel studerer vi kun to typer skjevhet. Likevel er eksempler på alle de nevnte skjevhetene inkludert i seksjon 2.

I dagligtale og offentlig diskurs benyttes begrepet skjevhet eller bias ofte for å beskrive at databaserte modeller forskjellsbehandler på sosialt eller etisk kritikkverdige måter. Selv om en slik begrepsbruk kan virke intuitiv, blander den ofte statistiske forhold i data og modell med normative vurderinger og sosiale og etiske betraktninger på utvikling og anvendelse.

På den ene siden er en tverrfaglig tilnærming til utfordringen nødvendig, kanskje særlig i en velferdskontekst, hvor både data og modell er en samkonstruksjon av sosiale, tekniske og rettslige forhold. Da kan modeller preges av skjevheter fra det underliggende datagrunnlaget, i selve modellen eller dens praktiske anvendelse. En god modellkritikk må dissekere slike systemer med verktøy fra ulike fagdisipliner. På den annen side kan utydelig begrepsbruk komplisere tverrfaglige diskusjoner og i verste fall maskere problematiske forhold. En tydeliggjøring av skjevhetsbegrepet er viktig for at nødvendige diskusjoner blir løftet i det offentlige rom, og for ettersyn, demokratisk kontroll og legitimitet i velferdssektoren.

Konklusjonen er at skjevhet er et mangefasettert konsept som må tas høyde for i alle deler av en datadrevet prosess. Det finnes ingen universelle løsninger eller retningslinjer for å avdekke eller håndtere skjevhet. Likevel gir vi følgende anbefalinger for å minimere risikoen for utilsiktet produksjonssetting av skjeve modeller:

- Tverrfaglig samarbeid for å gjøre utviklingsteamet bedre rustet til å evaluere datagrunnlag og modell.
- Vurder skadepotensial og hvilke konsekvenser av skjevhet som er uakseptable, for lettere å prioritere kontroll og ettergang av skjevhet.
- Gjør evaluering av skjevheter til en naturlig del av modellutviklingen for å avdekke problemer tidligere og treffe nødvendige tiltak.
- Publiser evalueringer av skjevheter for bredere innsyn og debatt rundt modellene.
- Monitorer modellene kontinuerlig for å sjekke at antagelsene som inngikk ved utvikling av modellen, fremdeles står ved lag.

Referanser

J. K. A. Aronson, D. Badenoch, A. Banerjee, C. Bankhead, J. A. Brassey, I. Chalmers, R. Davis, C. Friedemann-Smith, C. Heneghan, J. Lach, K. Mahtani, M. McCall, E. McFadden, D. Nunan, J. O'Sullivan, I. Onakpoya, A. Pluddemann, G. Richards, E. A. Spencer, & A. Turk. Catalogue of bias. https://catalogofbias.org/biases/, 2021. Lest: 2021-08-25.

S. Barocas, M. Hardt, & A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

K. Belitz & P. E. Stackelberg. Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling & Software*, 139:105006, 2021. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2021.105006.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, & Y. Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, & A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

J. Buolamwini & T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of Proceedings of *Machine Learning Research*, pages 77–91. PMLR, 2018. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Centre for Data Ethics and Innovation. Review into bias in algorithmic decision-making. https://www.gov.uk/government/publications/ cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making# contents, 2020. Lest: 2021-08-31.

T. Chen & C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, 2016. ISBN 978-1-4503-4232-2. doi: http://doi.acm.org/10.1145/2939672.2939785.

D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, & G. E. Dahl. On empirical comparisons of optimizers for deep learning, 2019.

C. Cortes, M. Mohri, M. Riley, & A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008. doi: https://doi.org/10.1007/978-3-540-87987-9_8.

I. Covert, S. Lundberg, & S.-I. Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22 (209):1–90, 2021. URL http://jmlr.org/papers/v22/20-1316.html.

D. R. Cox. *Principles of statistical inference*. Cambridge university press, 2006.

Datatilsynet. Varsler vedtak om at IB-karakterene må rettes.
https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-2020/varsler-vedtak-om-at-ib-karakterene-ma-rettes/, 2020. Lest: 2021- 08-31.

P. Dhrymes. *Misspecification Analysis and Errors in Variables*, pages 293–352. Springer International Publishing, 2017. ISBN 978-3-319-65916-9. doi: https://doi.org/10.1007/978-3-319-65916-9_5.

European Commission. Proposal for a regulation laying down harmonised rules on Artificial Intelligence, 4 2021. URL https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

E. L. Feige & H. W. Watts. An investigation of the consequences of partial aggregation of micro-economic data. *Econometrica*, 40(2):343–360, 1972. ISSN 00129682, 14680262. doi: https://doi.org/10.2307/1909411.

A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019. ISBN 9781492032649.

S. C. Geyik, S. Ambler, & K. Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2221–2231, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: https://doi.org/10.1145/3292500.3330691.

M. A. Gianfrancesco, S. Tamang, J. Yazdany, & G. Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, 11 2018. doi: https://doi.org/10.1001/jamainternmed.2018.3763.

P. Gijsbers, E. LeDell, J. Thomas, S. Poirier, B. Bischl, & J. Vanschoren. An open source automl benchmark, 2019.

U. S. Goth, H. L. Hammer, & B. Claussen. Utilization of Norway's emergency wards: the second 5 years after the introduction of the patient list system. *International journal of environmental research and public health*, 11(3):3375–3386, 2014. doi: https://doi.org/10.3390/ijerph110303375.

S. Grønmo. Bias i forskning. https://snl.no/bias_i_forskning, 2020.

M. Hammersley & R. Gomm. Bias in social research. *Sociological Research Online*, 2(1):7–19, 1997. doi: https://doi.org/10.5153/sro.55.

M. Hardt, E. Price, & N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett, editors, *Advances in neural information processing systems*, volume 29, pages 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. ISSN 00129682, 14680262. doi: https://doi.org/10.2307/1912352.

T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, 1995. doi: https://doi.org/10.1109/ICDAR.1995.598994.

A. E. Hoerl & R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: https://doi.org/10.1080/00401706.1970.10488634.

J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, & A. Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

F. Huettner & M. Sunder. Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Electronic Journal of Statistics*, 6: 1239–1250, 2012. doi: https://doi.org/10.1214/12-EJS710.

Independent High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustwothy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf, 2019. Lest: 2021-08-31.

G. James, D. Witten, T. Hastie, & R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

J. Kleinberg, S. Mullainathan, & M. Raghavan. Inherent tradeoffs in the fair determination of risk scores, 2016.

R. Klinkenberg. Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis*, 8(3):281–300, 2004. ISSN 1088-467X. doi: https://doi.org/10.3233/IDA-2004-8305.

Kommunal- og moderniseringsdepartementet. Nasjonal strategi for kunstig intelligens. https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/, 2020. Lest: 2021-08-31.

A. Liu & B. D. Ziebart. Robust classification under sample selection bias. In *NIPS*, pages 37–45, 2014.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, & S.-I. Lee. Explainable ai for trees: From local explanations to global understanding, 2019.

J. Mary, C. Calauzènes, & N. El Karoui. Fairness-aware learning for continuous attributes and treatments. In K. Chaudhuri & R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391. PMLR, 09-15 Jun 2019. URL http://proceedings.mlr.press/v97/mary19a.html.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, & A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. doi: https://doi.org/10.1145/3457607.

T. M. Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers University, 1980.

M. Ordu, E. Demir, C. Tofallis, & M. M. Gunal. A novel healthcare resource allocation decision support tool: A forecasting-simulation-optimization approach. *Journal of the Operational Research Society*, 72(3): 485–500, 2021. doi: https://doi.org/10.1080/01605682.2019.1700186.

L. A. Page & M. Henderson. Appraising the evidence: what is measurement bias? *Evidence-Based Mental Health*, 11(2):36–37, 2008. ISSN 1362-0347. doi: http://dx.doi.org/10.1136/ebmh.11.2.36.

J. C. Schlimmer & R. H. Granger. Incremental learning from noisy data. *Machine learning*, 1(3):317–354, 1986. doi: https://doi.org/10.1007/BF00116895.

L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, 1953.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

Språkrådet. På godt norsk – avløserord. https://www.sprakradet.no/sprakhjelp/Skriverad/Avloeysarord/#B, 2021.

R. Srinivasan & A. Chander. Biases in ai systems: A survey for practitioners. *Queue*, 19(2):45–64, apr 2021. ISSN 1542-7730. doi: https: //doi.org/10.1145/3466132.3466134.

A. Staff. Bias. https://www.forskningsetikk.no/ressurser/fbib/uavhengighet/bias/, 2015.

H. Suresh & J. Guttag. Understanding potential sources of harm throughout the machine learning life cycle. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, 8 2021. doi: https://doi.org/10.21428/2c646de5.c16a07bb.

N. Syed, H. Liu, & K. Sung. Handling concept drifts in incremental learning with support vector machines. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, KDD '99, pages 317–321. ACM, 1999. ISBN 1581131437. doi: https://doi.org/10.1145/312129.312267.

S. Verma & J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1– 7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: https://doi.org/10.1145/3194770.3194776.

H. P. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14:65–72, 1985. doi: https://doi.org/10.1007/BF01769885.

I. Žliobaitė, M. Pechenizkiy, & J. Gama. *An overview of concept drift applications*, pages 91–114. Springer International Publishing, Cham, 2016. ISBN 978-3-319-26989-4. doi: https://doi.org/10.1007/978-3-319-26989-4_4.

A Vedlegg

A.1 Begreper og oversikt over terminologi

I denne artikkelen bruker vi gjennomgående *skjevhet* som oversettelse for det engelske begrepet *bias*, men andre alternativer benyttes også i den norske faglitteraturen eller anbefales av norske institusjoner. Ved siden av *skjevhet* går *bias* oftest igjen i den norske faglitteraturen. Språkrådet anbefaler *slagside* som avløserord for *bias* (Språkrådet, 2021), men begrepet har foreløpig fått lite fotfeste i den norske faglitteraturen. Tidsskrift for Den norske legeforening anbefaler å bruke *skjevhet* (Staff, 2015). Forskningsetikk.no viser til Språkrådet fra Tidsskrift for Den norske legeforening, men virker selv å foretrekke *bias* (Staff, 2015).

Som vist i seksjon 2, finnes flere typer skjevhet uten etablerte norske betegnelser. Tabell S1 gir en oversikt over disse begrepene og de foreslåtte oversettelsene i denne artikkelen.

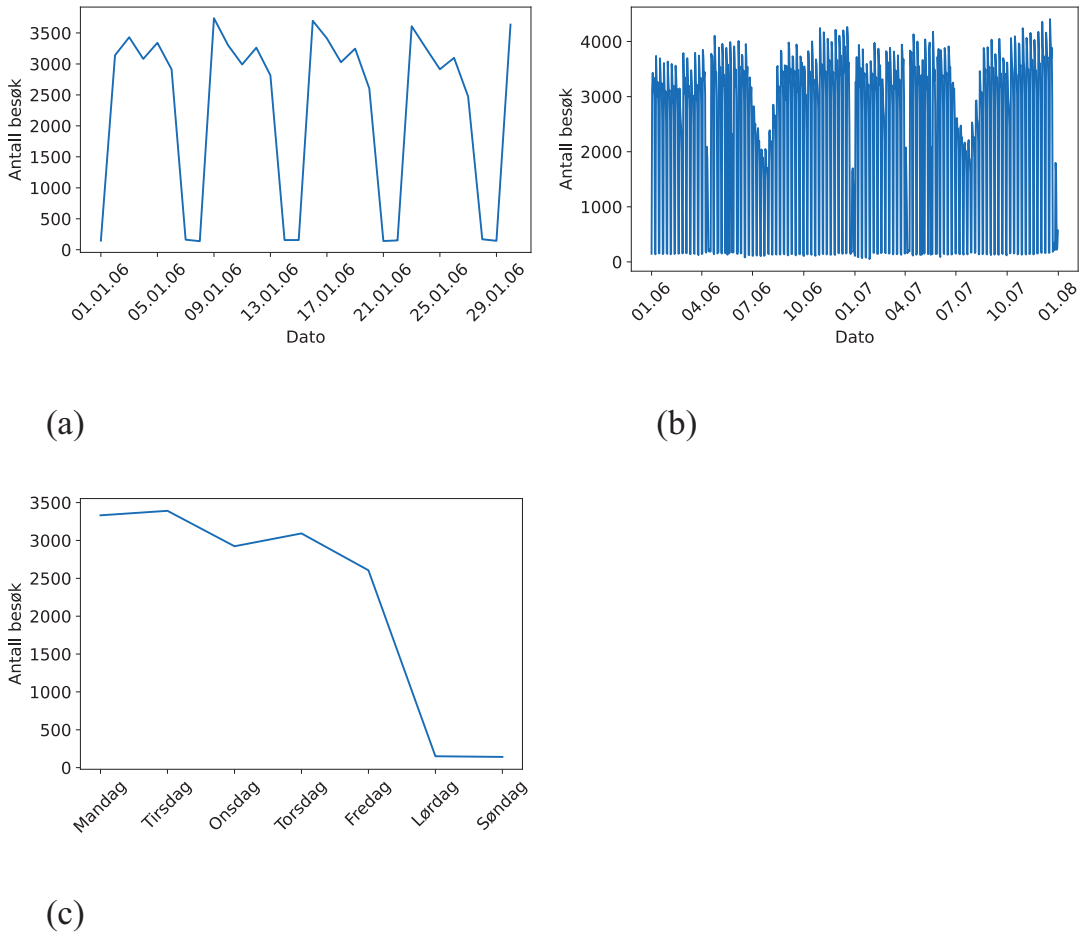
Tabell S1: Engelske begreper og norske oversettelser foreslått i denne artikkelen.

Engelsk begrep	**Foreslått oversettelse**
Bias	Skjevhet
Historical bias	Historisk skjevhet
Selection bias	Utvalgsskjevhet
Registration bias	Registreringsskjevhet
Measurement bias	Måleskjevhet
Representation bias	Representasjonsskjevhet
Learning bias	Læringsskjevhet
Evaluation bias	Referanseskjevhet
Aggregation bias	Aggregeringsskjevhet
Concept drift	Konseptdrift

A.2 Data- og problembeskrivelse

Analysene er basert på et avidentifisert registerdatasett fra NAV som inneholder samlede besøk til fastlege og legevakt for voksne i Norge, i 20 kommuner fra hele landet, i årene 2006–2007 og 2009–2010. Datasettet ble først brukt til å analysere legevaktbruk etter innføringen av fastlegeordningen i 2001 (Goth et al., 2014). Vi teller først opp totalt antall legekonsultasjoner per uke og per kommune. Figur S1 viser totalt antall legekonsultasjoner i Oslo kommune over én måned (januar 2006) og hele det første datasettets periode (2006 og 2007). Figur S1c viser gjennomsnittlig antall legekonsultasjoner i Oslo kommune i 2006 per ukedag. Figurene S1c og S1a viser en tydelig trend gjennom uken med markant færre legekonsultasjoner i helgene, og figur S1b viser en nedgang i legebesøk i ferier.

Antall legebesøk per kommune og uke sorteres på pasientgruppene < 70 år og ≥ 70 år. Vi utvikler maskinlæringsmodeller som predikerer antall legekonsultasjoner for pasientgruppen ≥ 70 i en kommune den kommende uka basert på antall legekonsultasjoner i hver av de 20 kommunene de tre foregående ukene for den aktuelle pasientgruppen (≥ 70) samt pasientgruppen < 70. Motivasjonen for å inkludere informasjon fra andre kommuner er at de kan inneholde nyttig informasjon, som spredning av smittsomme sykdommer.



Figur S1: (a) Antall legekonsultasjoner i Oslo kommune i januar 2006 og (b) i hele 2006 og 2007. (c) Gjennomsnittlig antall legekonsultasjoner i 2006 for personer bosatt i Oslo kommune gruppert etter ukedag.

Fire ulike læringsalgoritmer av ulik kompleksitet benyttes:

1. *XGBoost* (Chen & Guestrin, 2016), som består av en sekvens med trebaserte modeller, der hver modell korrigerer prediksjonene til den foregående. XGBoost oversettes ikke til norsk fordi det er et egennavn.
2. *Random Forest* (Ho, 1995), som består av en tilfeldig initialisering av beslutningstrær, og som vi derfor betegner som tilfeldig-skog-modell.
3. *Ridge-regresjon* (Hoerl & Kennard, 1970), som er en regularisert form for lineær regresjon.
4. *Polynomisk regresjon* (James et al., 2013), som er et spesialtilfelle av multippel lineær regresjon, der de opprinnelige forklaringsvariablene opphøyes i heltallseksponenter og danner nye forklaringsvariabler som inkluderes i modellen.

Se for eksempel Géron (2019) for detaljer om de ulike læringsalgoritmene. Maskinlæringsmodellen som gir best resultater på testdatasettet til henholdsvis Oslo og Kragerø, benyttes videre for å illustrere effekten av skjevheter.

Maskinlæringsmodellene som benyttes i eksemplene, er relativt enkle i forhold til dyplæringsmodeller. Likevel er problemstillingene de belyser, modellagnostiske. Læringsalgorit-

mene er aktuelle for kvantitativ analyse av strukturerte data (Géron, 2019), noe som er en vanlig datatype i en velferdskontekst som et resultat av hvordan dataene samles inn.

A.3 Forklaringsmetoder

Forklaringsmetoder basert på det spillteoretiske løsningskonseptet Shapley-verdier (Shapley, 1953) er populære i maskinlæringslitteraturen. De baserer seg på å regne ut eller approksimere den såkalte Shapley-dekomposisjonen av bidragene de ulike variablene utgjør for modellen. Intuitivt kan Shapley-verdier ses på som den rettferdige andelen av den totale gevinsten i et lagspill som hver spiller som deltar i spillet, bør få. Shapley-dekomposisjonen har et solid teoretisk grunnlag (Young, 1985) og er basert på et sett aksiomer som fører til flere appellerende egenskaper (Huettner & Sunder, 2012). Blant annet vil variabler som bidrar like mye, få den samme Shapley-verdien, og variabler som ikke bidrar, får verdien 0. Videre kan Shapley-verdier beregnes for alle typer modeller inkludert mer avanserte maskinlæringsmodeller, der tradisjonelle forklaringsmetoder som regresjonskoeffisienter kommer til kort. Det er allerede en utstrakt bruk av forklaringsmetoder som er basert på Shapley-verdier, slik som SHAP (Lundberg et al., 2019) og SAGE (Covert et al., 2021). Å beregne eksakte Shapley-verdier er beregningstungt, da alle kombinasjonene av samtlige spillere (variabler) inkludert og ekskludert fra spillet (modellen) må tas med i beregningen. Metodene SHAP og SAGE løser dette ved hjelp av raske approksimasjoner, og begge er åpent tilgjengelige i form av Python- og R-biblioteker.

A.4 Korreksjonsmetoder for skjevhet

Ideelt sett trenes en maskinlæringsmodell på data som er representative for populasjonen den skal brukes på. Når dette ikke kan garanteres, kan deteksjons- og korreksjonsmetoder for skjevhet brukes. I dette avsnittet introduserer vi et utvalg slike metoder med fokus på skjevhet i forklaringsvariabler og konseptdrift.

Hvis fordelingen i populasjonen modellen skal brukes på, er kjent, kan modellen trenes på en tilsvarende fordeling. Alternativt, eller hvis feilestimering av den underrepresenterte gruppen har store negative konsekvenser, kan ulik vekting av klassene i treningssettet benyttes under utvikling. Dersom skjevheten ikke er kjent, kan en probabilistisk tilnærming med *kernel density estimator (KDE)* (Shimodaira, 2000) eller en *robust bias-aware* metode (Liu & Ziebart, 2014) benyttes til å estimere fordelingen. Sistnevnte metode estimerer fordelingen til testsettet og re-vekter de observerte verdiene i treningssettet deretter. Gruppering av observasjonene i testsettet kan også brukes for å estimere fordelingen ved at denne sammenliknes med fordelingen i treningssettet (Cortes et al., 2008). Empirisk distribusjonstilpasning sammenlikner distribusjonen til modellens prediksjoner med distribusjonen til de observerte verdiene og bruker regresjon til å korrigere modellens prediksjoner (Belitz & Stackelberg, 2021). Alternativt kan treningsdataene vektes basert på gjennomsnittsverdier i trenings- og testdataene (Huang et al., 2006). Det er også mulig å justere testdataene slik at fordelingen nærmer seg fordelingen til dataene modellen er trent på.

Problemer relatert til konseptdrift kan håndteres på ulike måter. En mye brukt tilnærming er å re-trene modellen med jevne mellomrom eller ved oppdagelse av konseptdrift (Žliobaitė et al., 2016). Avhengig av type maskinlæringsmodell kan man også vekte de nyeste observasjonene mer enn de eldre og slik oppmuntre modellen til å lære de nye konseptene (Klinkenberg, 2004). Alternativt kan en ny maskinlæringsmodell baseres på trender lært av den opprinnelige modellen og trenes på nye data (Syed et al., 1999). En ulempe med disse fremgangsmåtene er at det kan være ressurskrevende å re-trene modeller. Dersom modellen kun re-trenes på data som er samlet etter at konseptdriften inntraff, vil man også ha få observasjoner å trene den nye modellen på.

A.5 Nærmere beskrivelse av underrapportering

A.5.1 Oslo kommune

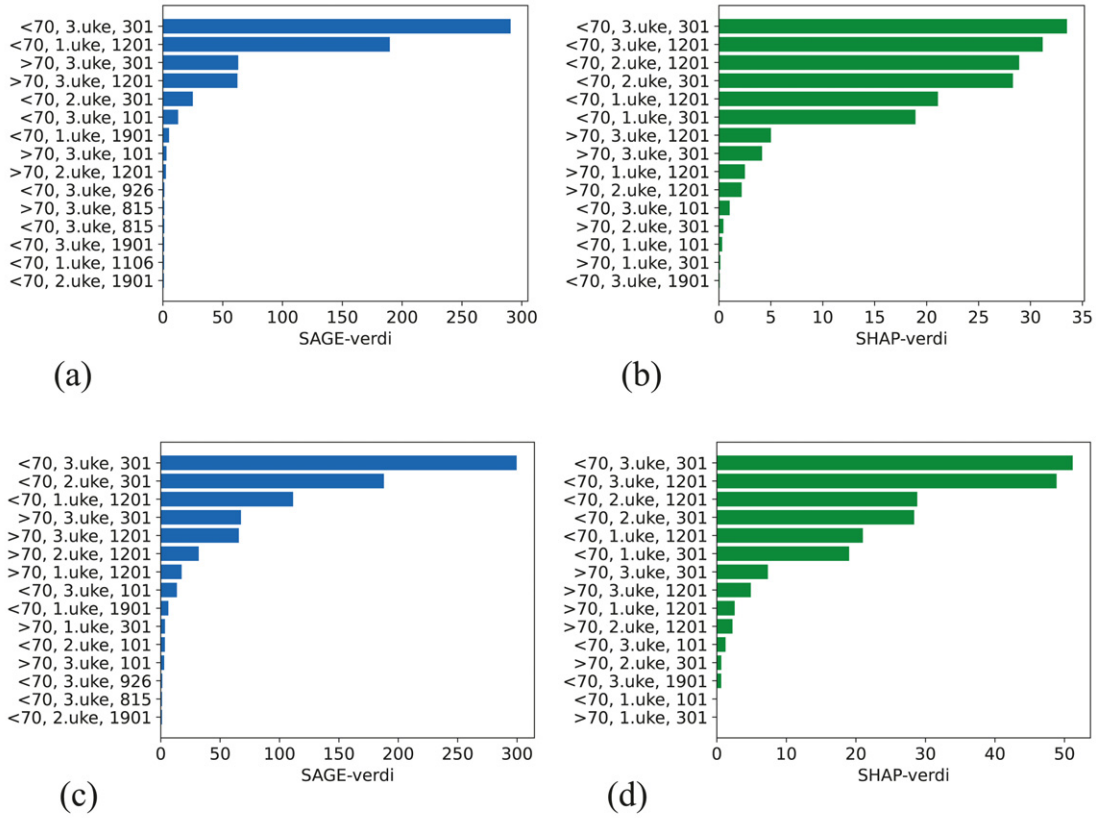
Vi trener flere maskinlæringsmodeller på data fra 2006, 2007 og 2009 til å predikere antall ukentlige besøk for personer på ≥ 70 år i Oslo kommune (301) og tester dem på data fra 2010. Siden Ridge-regresjon gir best resultater på testdatasettet uten skjevheter, med en RMSE på 750, bruker vi denne for prediksjoner i Oslo. Deretter predikerer vi antall ukentlige legekonsultasjoner for gruppen av personer på ≥ 70 år i Oslo kommune (301) for data fra 2010, med skjevheten i uken før prediksjon. Vi observerer at RMSE øker til 1,446.

A.5.2 Kragerø kommune

Modellen som predikerer antall legekonsultasjoner for personer ≥ 70 år i Kragerø kommune (kommunenummer 815), trenes og testes på samme måte som modellen i seksjon 3.2.1. Igjen benyttes Ridge-regresjon, som oppnår laveste RMSE (30) på testdatasettet.

A.6 Nærmere beskrivelse av korreksjon av underrapportering

Følgende fremgangsmåte benyttes for å korrigere for underrapportering: Vi antar at vi ikke vet hvilke variabler som har skjevhet, og korrigerer derfor alle forklaringsvariablene i testdataene. Dette gjøres ved hjelp av gjennomsnittsverdier fra treningsdataene; først benyttes gjennomsnittlig antall konsultasjoner for uke 1 i henholdsvis 2006, 2007 og 2009 til å justere konsultasjonstallene for uke 1 i 2010. Deretter predikerer modellen konsultasjonstall for påfølgende uke i datasettet der de foregående ukene har blitt korrigert. Jo lenger ut i året vi kommer, jo mer data kan korrigeres, og jo bedre bør modellens treffsikkerhet bli.



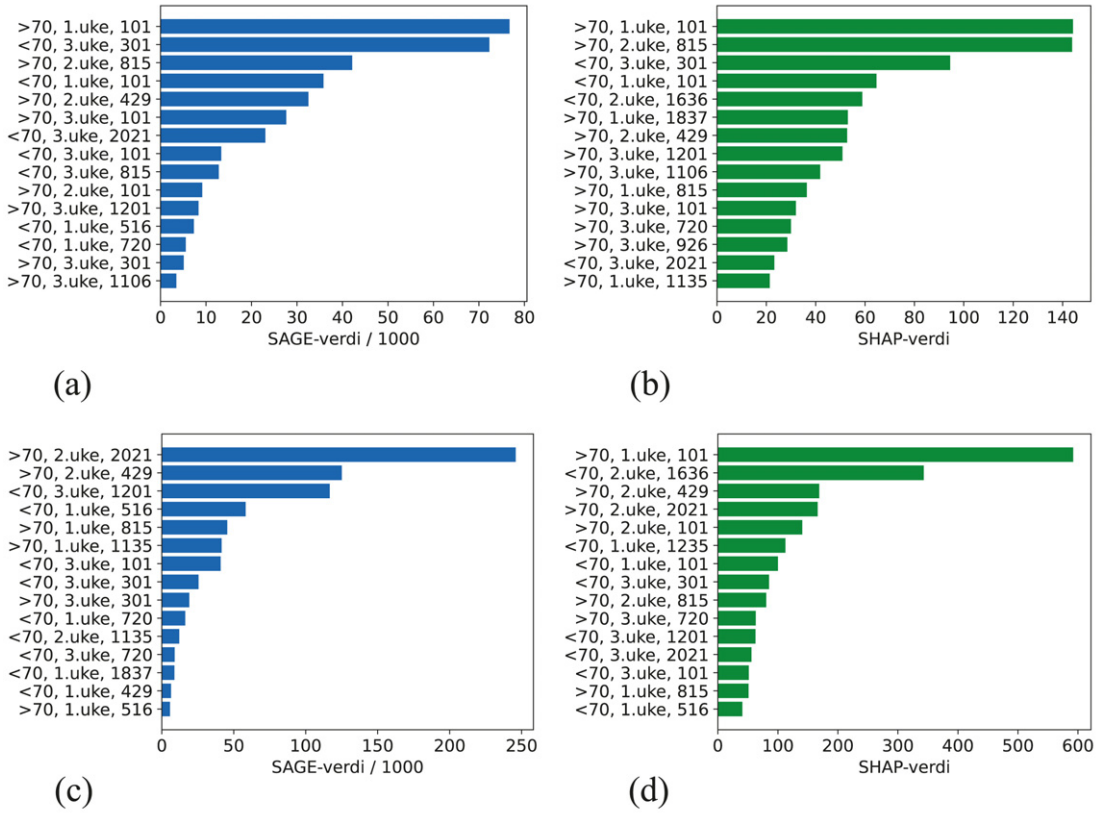
Figur S2: (a) SAGE-verdier og (b) SHAP-verdier for prediksjoner på personer med alder ≥ 70 år i Kragerø kommune i 2010. (c) Tilsvarende SAGE-verdier og (d) SHAP-verdier ved kunstig lave besøkstall for foregående uke.

A.7 Skjevhet grunnet endring over tid

To nye modeller trenes på data fra 2006 for å predikere antall legekonsultasjoner for personer ≥ 70 år i henholdsvis Oslo og Kragerø. Modellene testes først på data fra 2007. For å studere effekten av konseptdrift (seksjon 2.2) lar vi deretter modellene predikere antall legekonsultasjoner for 2010. Maskinlæringsalgoritmene som benyttes for å predikere antall legekonsultasjoner for Oslo og Kragerø kommune, er henholdsvis XGBoost og Ridge-regresjon, da disse oppnår lavest RMSE på testsettene.

A.7.1 Oslo kommune

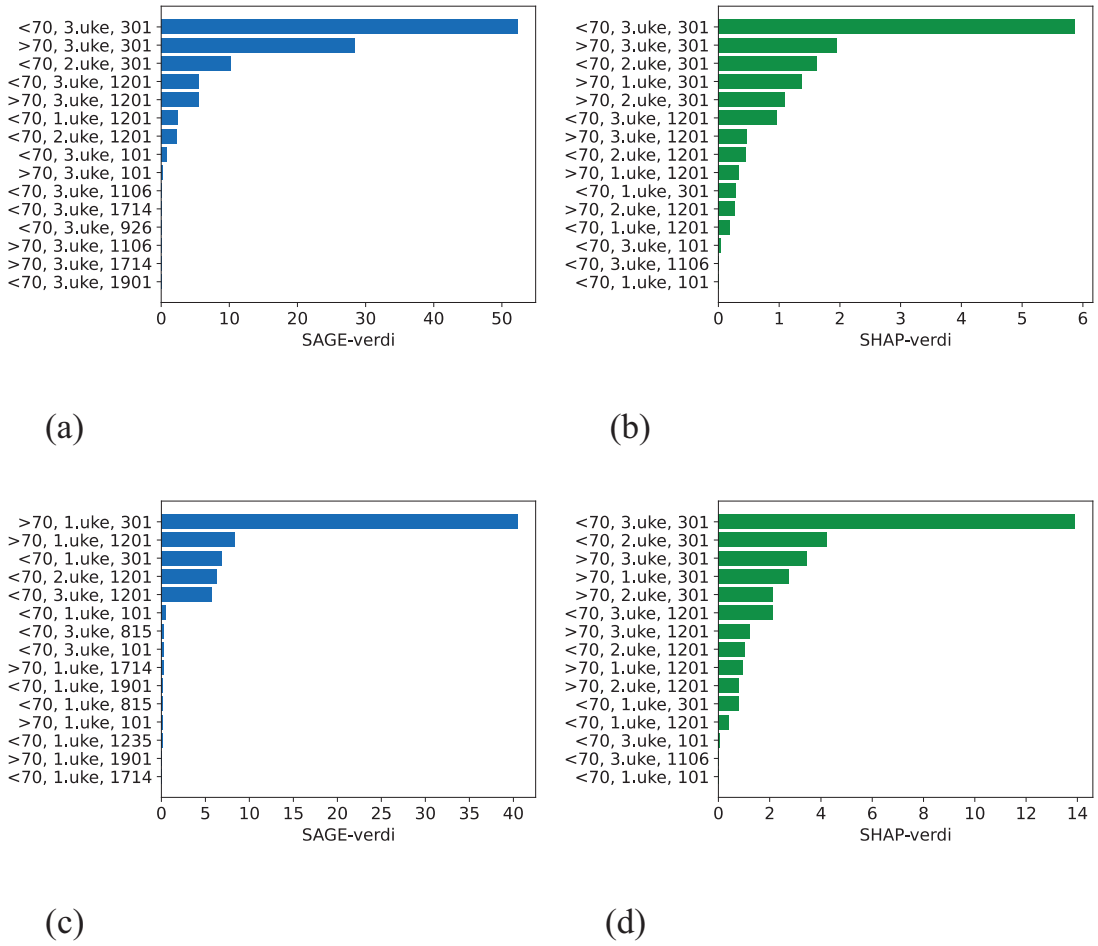
Modellens RMSE på testdatasettet fra 2007 er på 711, men denne øker til 2295 etter innføring av konseptdrift, altså ved prediksjoner på testdatasettet fra 2010. Dette skyldes sannsynligvis at relasjonen mellom forklaringsvariablene og den forklarte variabelen har endret seg mellom dataene fra 2007 og 2010. Ifølge SAGE- og SHAP-verdiene i figurene S3a og S3b er antall legekonsultasjoner for personer i Halden (101), Kragerø (815) og Oslo (301) viktigst for modellen før konseptdrift. Karasjok (2021), Lillesand (429) og Bergen (1201) er de viktigste kommunene etter innføring av konseptdrift ifølge SAGE-verdiene, mens Oslo (301), Meldal (1636) og Lillesand (429) er viktigst ifølge SHAP-verdiene, se figurene S3c og S3d. Årsaken til de observerte endringene er uviss og kan skyldes flere faktorer. Når modellen predikerer dårligere, vektlegger den forklaringsvariabler som er mindre åpenbart viktige for å predikere konsultasjoner i Oslo kommune.



Figur S3: (a) SAGE-verdier og (b) SHAP-verdier for prediksjoner på personer med alder ≥ 70 år i Oslo kommune i 2007. (c) Tilsvarende SAGE-verdier og (d) SHAP-verdier etter innføring av konseptdrift

A.7.2 Kragerø kommune

Tilsvarende forsøk som for Oslo gjentas for Kragerø. Ved innføring av konseptdrift øker RMSE fra 22 til 38. Dette gjenspeiler resultatene fra konseptdrift for Oslo. Ved å studere SAGE- og SHAP-verdiene (vedlagt figur S4a og figur S4b) ser vi at ukentlige konsultasjonstall for personer i Oslo (301) er de viktigste forklaringsvariablene før konseptdrift. Etter innføring av konseptdrift er tall for Bergen (1201) på andreplass ifølge SAGE-verdiene, men ikke ifølge SHAP-verdiene, se figur S4c og S4d. Absoluttverdiene til modellen er svært annerledes etter konseptdrift, og modellprestasjonen er redusert.

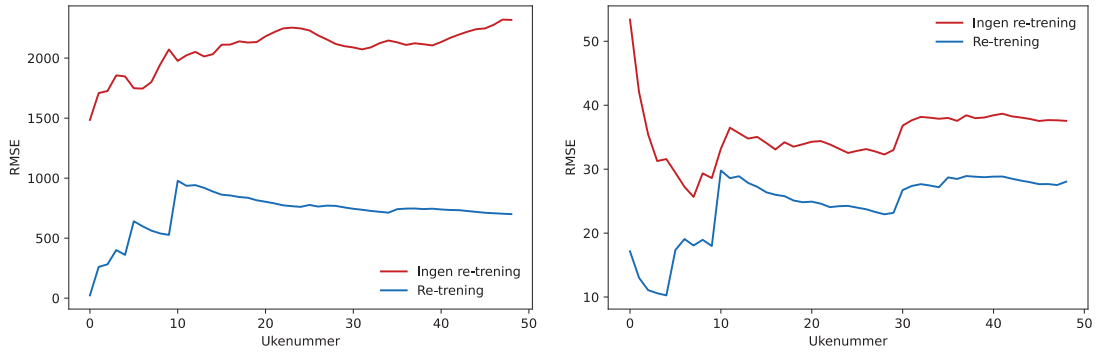


Figur S4: (a) SAGE-verdier og (b) SHAP-verdier for prediksjoner på personer med alder ≥ 70 år i Kragerø kommune i 2007. (c) Tilsvarende SAGE-verdier og (d) SHAP-verdier etter innføring av konseptdrift.

A.7.3 Korreksjon av endring over tid

Opprinnelig ble modellene for å illustrere konseptdrift trent på data fra 2006. Vi så at RMSE-verdiene økte da modellene predikerte på data fra 2010. Vi re-trener modellene på data fra 2009 for å se om dette gir bedre prediksjoner for ukentlige besøkstall i 2010. Modellenes prediksjoner uke for uke i 2010 både før og etter re-trening på nyere data vises i S5a og S5b. RMSE-verdiene er i begge tilfeller lavere etter at modellene er re-trent (blå kurver). En bratt økning i prediksjonsfeil observeres ved uke 11. De registrerte konsultasjonstallene for denne uken er markant lavere for både Oslo og Kragerø, noe som igjen kan skyldes påsken. RMSE-

verdiene for prediksjoner på data fra 2010 etter re-trening av modellen synker fra 2295 til 782 for Oslo, og fra 38 til 30 for Kragerø. Selv om prediksjonene ikke er optimale, er den re-trente modellen mer treffsikker.



Figur S5: Prediksjonsfeil for konsultasjonstall for personer med alder ≥ 70 i Oslo (a) og Kragerø (b). Blå kurver tilsvarer prediksjoner etter at modellen er re-trent på nye data og rød kurver tilsvarer prediksjoner gjort av modellen uten re-trening. RMSE: Root Mean Squared Error

A.8 Skjevhet og rettferdighet

Selv om skjevhet og rettferdighet er ulike statistiske begreper, kan de i praksis ofte samvirke og være relevante samtidig:

- Dersom en maskinlæringsmodell trenes opp på historiske data der en minoritetsgruppe er underrepresentert i registrene fordi den ikke har hatt like god tilgang på helsetjenester som majoriteten, kan maskinlæringsmodellen systematisk underpredikere tjenestebehovet for minoritetsgruppen. Dette vil både falle inn under rettferdighetsbegrepet og være en form for skjevhet, nemlig historisk skjevhet og utvalgsskjevhet.
- Når korrigerende tiltak påføres modellen for å imøtekomme utfallskrav, kan dette i praksis medføre en omdefinering av optimaliseringsproblemet og tapsfunksjonen til modellen. I så måte påfører rettferdighetskravet modellen en skjevhet mot et ønskemål som typisk er normativt motivert, og som kan utelukke konkurrerende ønskemål.
- Valget av hvilke egenskaper som skal tilfredsstille betinget uavhengighet, og hvilke som ikke behøver å gjøre det, er bestemmende for korreksjon og medfører også en form for modellskjevhet.

Til tross for at det finnes flere ulike statistiske mål på modellutfall som kan knyttes til ulike moralteoretiske ideer om rettferdighet, samsvarer de ikke nødvendigvis entydig med begrepet slik det benyttes i dagligtale eller i juridisk forstand. En maskinlæringsmodell kan eksempelvis godt utvise betinget avhengighet til en gitt gruppe, uten at dette betegnes som urettferdig: En hypotetisk prioriteringsmodell for NAV-veiledere som gir en prioritert gruppe et relativt fortrinn, kan tenkes å være et tilfelle av rettmessig forskjellsbehandling. Noen ganger er det nettopp en betinget avhengighet som sikrer et ønskelig utfall: ved å sørge

for variasjon betinget i for eksempel kjønn og alder kan kandidatsøk gjøres mer varierte, til tross for at noen grupper er statistisk underrepresentert i datagrunnlaget (Geyik et al., 2019).

Det bør også merkes at rettferdighetsbegrepet i alminnelighet ikke bare omfatter utfall, men også prosess, innretning og anvendelse. I juridisk forstand skilles det eksempelvis gjerne mellom *prosessuell* og *substansiell* rettferdighet, hvor rettferdighetsmål knyttet til utfallsskjevheter fra maskinlæringsmodeller gjerne adresserer sistnevnte. En modell som er vurdert som rettferdig i sine utfall, kan samtidig være prosessuelt urettferdig hvis for eksempel datainnsamlingen som ligger til grunn for modelltreningen blir vurdert som skjev, urettferdig eller ulovlig.