Kursus i algoritmiske grundprincipper og etik i AI: fra teori til praksis

Værktøjer til synkrone sessioner

CU5 | Casestudier og projekter

Formativ vurdering med åbne svar

Projektnummer | 2022-1-ES01-KA220-HED-000085257

Formativ vurdering med åbne svar

Formativ vurdering med åbne svar

	Beskrivelse	Kommentarer
Beskrivelse af opgaven	Introducerer de studerende til IBM's Fairness 360-værktøj, der har til formål at give indsigt i potentielle bias inden for demografi, hvordan det måles og strategier til afhjælpning, hvis der identificeres bias. Vi bruger Compas Propublica-applikationen som et eksempel.	
Beskrivelse af, hvordan opgaven skal udføres	AI Fairness 360 - Demo (ibm.com) Vælg Compas (ProPublica recidiv) som demoværktøj. Svar på følgende spørgsmål: 1) Til hvilket formål bruges Compas-algoritmen? 2) Hvilken slags bias har man opdaget i værktøjet? 3) Hvad er de beskyttede attributter? 4) Hvilke statistiske målinger viser bias i dette tilfælde for begge attributter? 5) Hvor meget ændrer den gennemsnitlige oddsforskel sig i tilfælde af "reweighing algorithm" i begge attributter? 6) Hvorfor er *reweighing* nødvendig, før applikationen tages i brug?	Flere detaljer og instruktioner findes på næste side.
Anslået tid til at udføre opgaven	40-60 minutter.	
Forslag til kilder til at udføre opgaven	AI Fairness 360-værktøjet giver alle svarene, bortset fra nummer 2 og 6, hvor du frit kan søge på internettet efter svaret.	
Detaljeret beskrivelse af, hvordan opgaven skal udføres	Underviseren bør forklare, hvor opgaven skal afleveres (f.eks. via e-mail, i en bestemt mappe, der tidligere er delt med den studerende, i et område, der er oprettet i kursusstrukturen på e-læringsplatformen ...).	
Information om deadline for aflevering af opgaven	Underviseren bør fastsætte en deadline for aflevering af denne opgave (vær opmærksom på kursets struktur med hensyn til asynkront arbejde og synkrone sessioner).	Oplys datoen i introduktionssessionen.
Kontaktoplysninger eller hvordan man afklarer tvivlsspørgsmål	Læreren skal angive en kontaktform.	Det kan være en e-mailadresse, et telefonnummer eller lign.

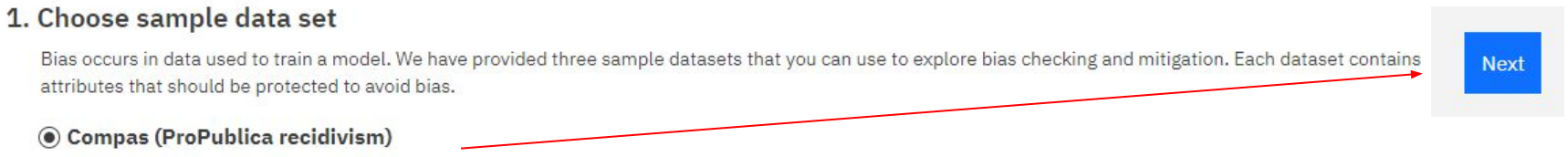
Formativ vurdering med åbne svar
Instruktioner

IBM's open source-værktøj AI Fairness 360 kan hjælpe dig med at undersøge, rapportere og afbøde diskrimination og bias i maskinlæringsmodeller i hele AI-applikationens livscyklus. Se hjemmesiden for detaljerede oplysninger: https://aif360.res.ibm.com/

I denne opgave undersøger vi en case i webdemoen.
Åbn demoværktøjet via dette link: AI Fairness 360 - Demo (ibm.com)

Her bruger vi Compas (ProPublica recidiv) som eksempel.

1) Hvad bruges Compas-algoritmen til? ⓘ Det er nævnt direkte i værktøjet.
2) Hvilken slags bias har man fundet i Compas Probulica-løsningen? ⓘ Søg på internettet efter et svar.
3) Hvad er de beskyttede attributter i løsningen? ⓘ Det kan ses direkte i værktøjet.
4) Hvilke statistiske målinger viser bias i dette tilfælde for begge attributter? ⓘ Vælg værktøjet ved at klikke på det, og vælg næste -> du kan se svaret på næste side.



5) Hvor meget ændres den gennemsnitlige oddsforskel i tilfælde af "reweighing algorithm" i begge attributter? ⓘ Du kan se svaret på dette, når du går til næste side og trykker på indstillingen "reweighing".
6) Hvorfor er *reweighing* nødvendig, før applikationen tages i brug? ⓘ For at finde svaret kan du f.eks. se dette link: Reweighting: Refining AI with Precision and Efficiency | by maxine | Medium

Formativ vurdering med åbne svar

Korrekte svar til den formative vurdering

Du kan indsamle svarene på den måde, der passer dig bedst. Svarene kan findes i selve værktøjet for alle spørgsmål undtagen 2 og 6.

1. **Hvad bruges Compas-algoritmen til?**

Strafudmåling i strafferetten.

2. Hvilken slags bias har man fundet i værktøjet?

Sorte tiltalte blev ofte anset for at have en højere risiko for tilbagefald til kriminalitet, end de faktisk havde.
Hvide tiltalte blev ofte anset for at være mindre tilbøjelige til at begå kriminalitet, end de var.
Analysen viste også, at selv når man kontrollerede for tidligere forbrydelser, fremtidig sandsynlighed for tilbagefald til kriminalitet, alder og køn, var der 45 procent større sandsynlighed for, at sorte tiltalte fik en højere risikoscore end hvide tiltalte.

Sorte tiltalte var også dobbelt så tilbøjelige som hvide tiltalte til at blive fejlklassificeret som havende en højere risiko for tilbagefald til voldelige handlinger (*violent recidivism*). Analysen af voldeligt recidiv viste også, at selv når man kontrollerede for tidligere forbrydelser, fremtidigt recidiv, alder og køn, var der 77 procent større sandsynlighed for, at sorte tiltalte fik en højere risikoscore end hvide tiltalte. (Kilde: How We Analyzed the COMPAS Recidivism Algorithm — ProPublica)

3. Hvad er de beskyttede attributter?

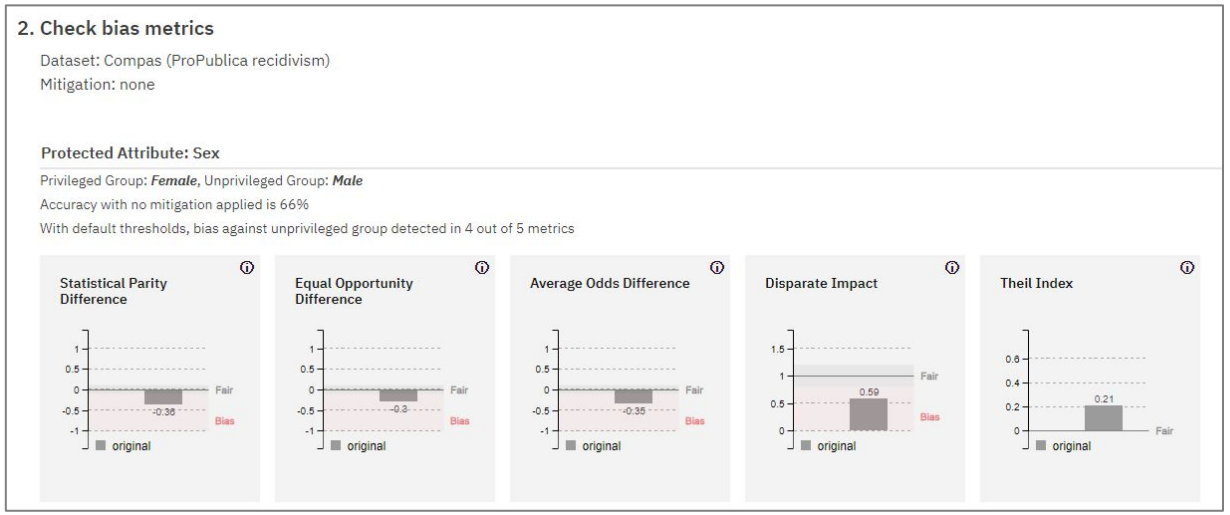
(Svaret kan ses direkte i værktøjet)
Køn, privilegeret: ***Kvinde***, ikke-privilegeret: ***Mand***
Race, privilegeret: ***Kaukasisk***, ikke-privilegeret: ***Ikke kaukasisk***

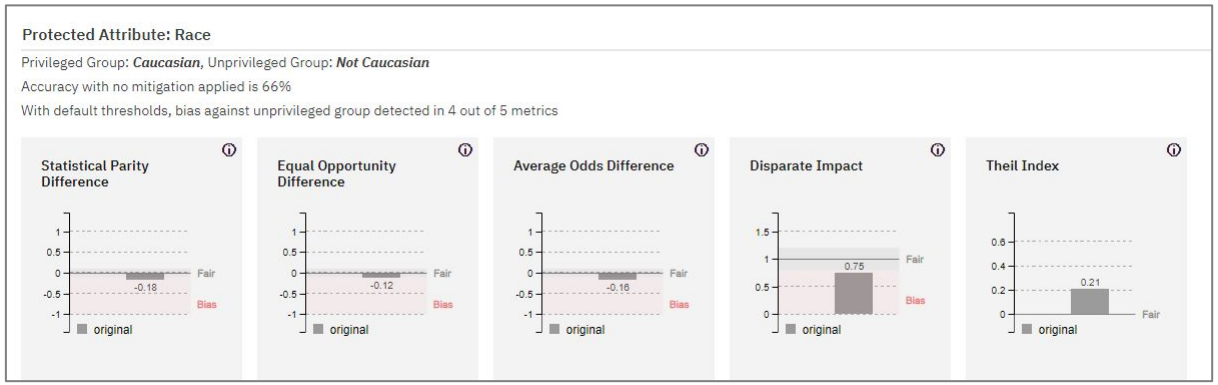
Formativ vurdering med åbne svar

Korrekte svar til den formative vurdering

4. Hvilke statistiske målinger viser bias i dette tilfælde for begge attributter?

Svaret fremgår af billederne nedenfor (fra værktøjet). For begge attributter vises der alle steder bias, dvs. vedr. *statistical parity difference, equal opportunity difference, average odds difference* og *disparate impact*.



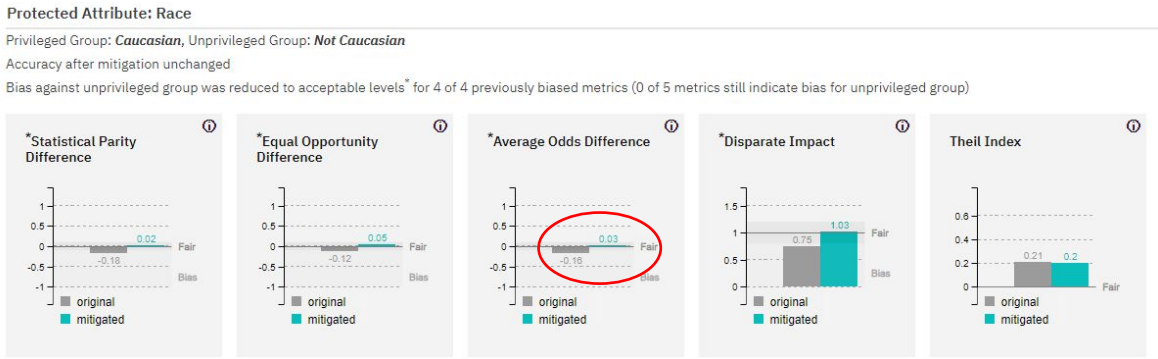


Formativ vurdering med åbne svar

Korrekte svar til den formative vurdering

5. Hvor meget ændrer den gennemsnitlige oddsforskel sig i tilfælde af "reweighing algorithm" i begge attributter?

Køn: 0,33 enheder
Race: 0,19 enheder



6) Hvorfor er *reweighing* nødvendig, før applikationen tages i brug?

Se linket her [Reweighting: Refining AI with Precision and Efficiency | by maxine | Medium]

Kort sagt: Det gør det muligt at rekalibrere algoritmen uden at bruge flere data.

TAK

Projektnummer: 2022-1-ES01-KA220-HED-000085257

Medfinansieret af Den Europæiske Union

Finansieret af Den Europæiske Union. Synspunkter og holdninger, der kommer til udtryk, er udelukkende forfatterens/forfatternes og er ikke nødvendigvis udtryk for Den Europæiske Unions eller Det Europæiske Forvaltningsorgan for Uddannelse og Kulturs (EACEA) officielle holdning. Hverken den Europæiske Union eller EACEA kan holdes ansvarlig herfor.