

Bazele algoritmice și etica în IA: de la teorie la practică

Set de instrumente pentru sesiuni
sincrone

**CU4 | Corectitudinea datelor și
prejudecățile în inteligența artificială**
Suport pentru diapozitive PowerPoint

INDEX

- INTRODUCERE - 3
- INTRODUCERE ÎN CORECTITUDINEA ȘI PĂRTINIREA DATELOR - 7
- TIPURI DE PREJUDECĂȚI ÎN SISTEMELE AI - 20
- METODE DE IDENTIFICARE ȘI MĂSURARE A PREJUDECĂȚILOR DIN DATE - 34
- STRATEGII PENTRU ABORDAREA ȘI ATENUAREA PREJUDECĂȚILOR - 41
- ÎNȚELEGEREA PARAMETRILOR DE MĂSURARE A CORECTITUDINII - 47
- ECHITATE ÎN COLECTAREA ȘI PRETRATAREA DATELOR - 58
- TENDINȚE VIITOARE ȘI PROBLEME EMERGENTE - 67
- CONCLUZIE - 76

INTRODUCERE



Sursa imaginii | Generat de DALL-E

Sugestie

Începeți cu un **sondaj sau o întrebare interactivă** pentru a evalua înțelegerea inițială sau experiențele personale ale participanților cu privire la corectitudinea și părtinirea AI. Acest lucru poate face introducerea mai atractivă și poate pregăti terenul pentru a explica de ce subiectul este relevant.

- Exemplu: **Care este gradul dvs. de familiaritate cu corectitudinea și prejudecățile în IA?**
 - Nu sunt deloc familiarizat cu ea.
 - Am auzit despre asta, dar nu cunosc detaliile.
 - Am înțeles elementele de bază, dar aș dori să știu mai multe.
 - Sunt destul de bine informat și am lucrat cu sisteme AI care iau în considerare corectitudinea și părtinirea.

ÎN ACEASTĂ UNITATE DE COMPETENȚĂ VEȚI GĂSI URMĂTOARELE SUBIECTE:

- Principiile de bază ale corectitudinii datelor și prejudecăților în inteligența artificială.
- Diverse prejudecăți care pot influența sistemele AI.
- Cum să identificați și să măsurați prejudecățile din date.
- Tehnici de reducere a prejudecăților și de creștere a corectitudinii.
- Metrici pentru evaluarea corectitudinii AI.
- Cele mai bune practici pentru colectarea și prelucrarea corectă a datelor.
- Tendințe emergente în corectitudinea IA. corectitudinea și prejudecățile în IA.

LA SFÂRȘITUL UNITĂȚII DE COMPETENȚĂ, AR TREBUI SĂ FIȚI CAPABIL SĂ:

- Explicați elementele fundamentale ale corectitudinii datelor și semnificația prejudecăților în inteligența artificială.
- Identificați diferitele forme de prejudecăți care pot afecta sistemele AI.
- Evaluați sistemele de inteligență artificială utilizând parametri de corectitudine.
- Urmați cele mai bune practici de colectare și preprocesare a datelor pentru a promova echitatea.

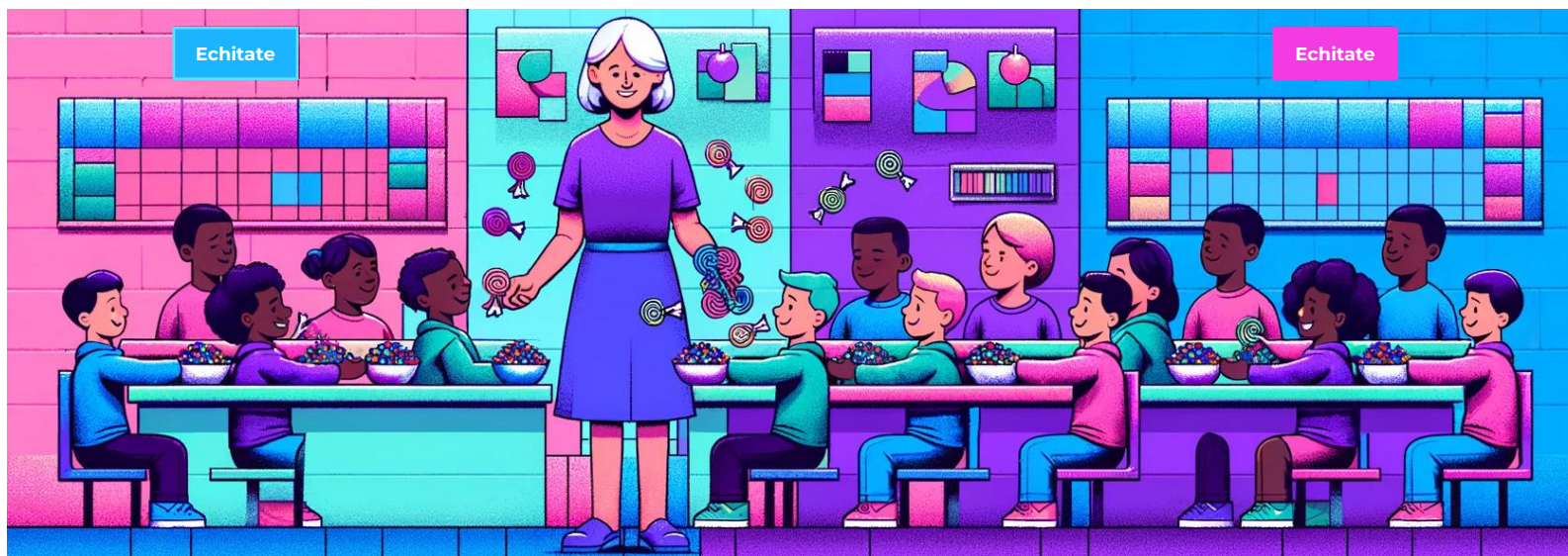
INTRODUCERE ÎN CORECTITUDINEA ȘI PĂRTINIREA DATELOR



Sursa imaginii | Generat de DALL-E

Introducere în corectitudinea și părtinirea datelor

Echitate și părtinire - analogia cu bomboanele



Sursa imaginii | Generat de DALL-E

Descoperiți modul în care corectitudinea și prejudecățile din sistemele de inteligență artificială pot influența alegerile noastre de zi cu zi, la fel cum regulile unui profesor pentru distribuirea bomboanelor pot influența experiența din clasă.

Haideți să dezvăluim conceptul de capital AI - pe rând!

Introducere în corectitudinea și părtinirea datelor

Echitate și părtinire - analogia cu bomboanele

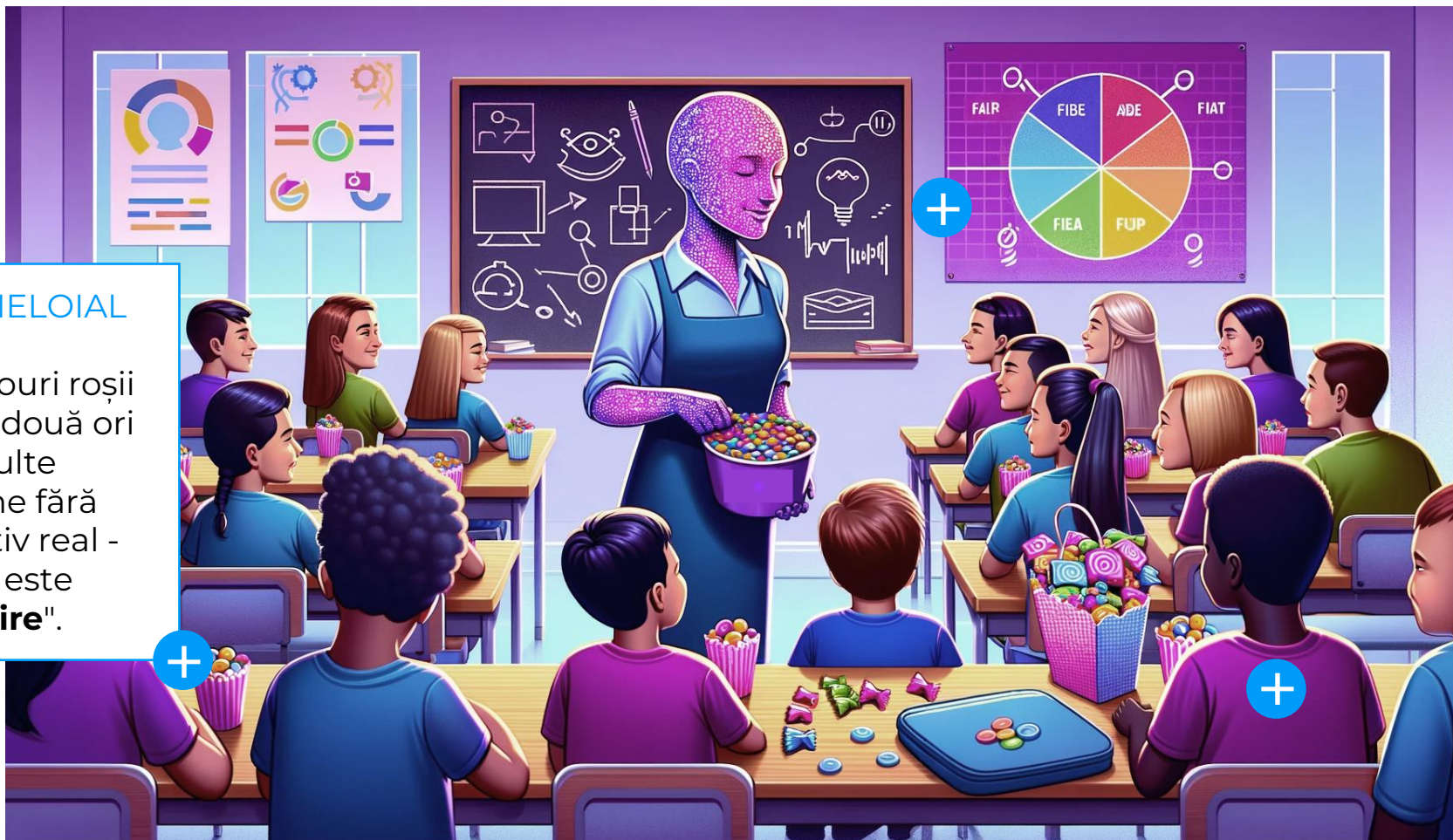
**Imaginează-ți o clasă în care șansele tale de a primi o bomboană în plus
depinde de ceea ce porți!**

Introducere în corectitudinea și părtinirea datelor

Echitate și părtinire - analogia cu bomboanele

AVANTAJ NELOIAL

Elevii în tricouri roșii primesc de două ori mai multe bomboane fără niciun motiv real - aceasta este "**părtinire**".



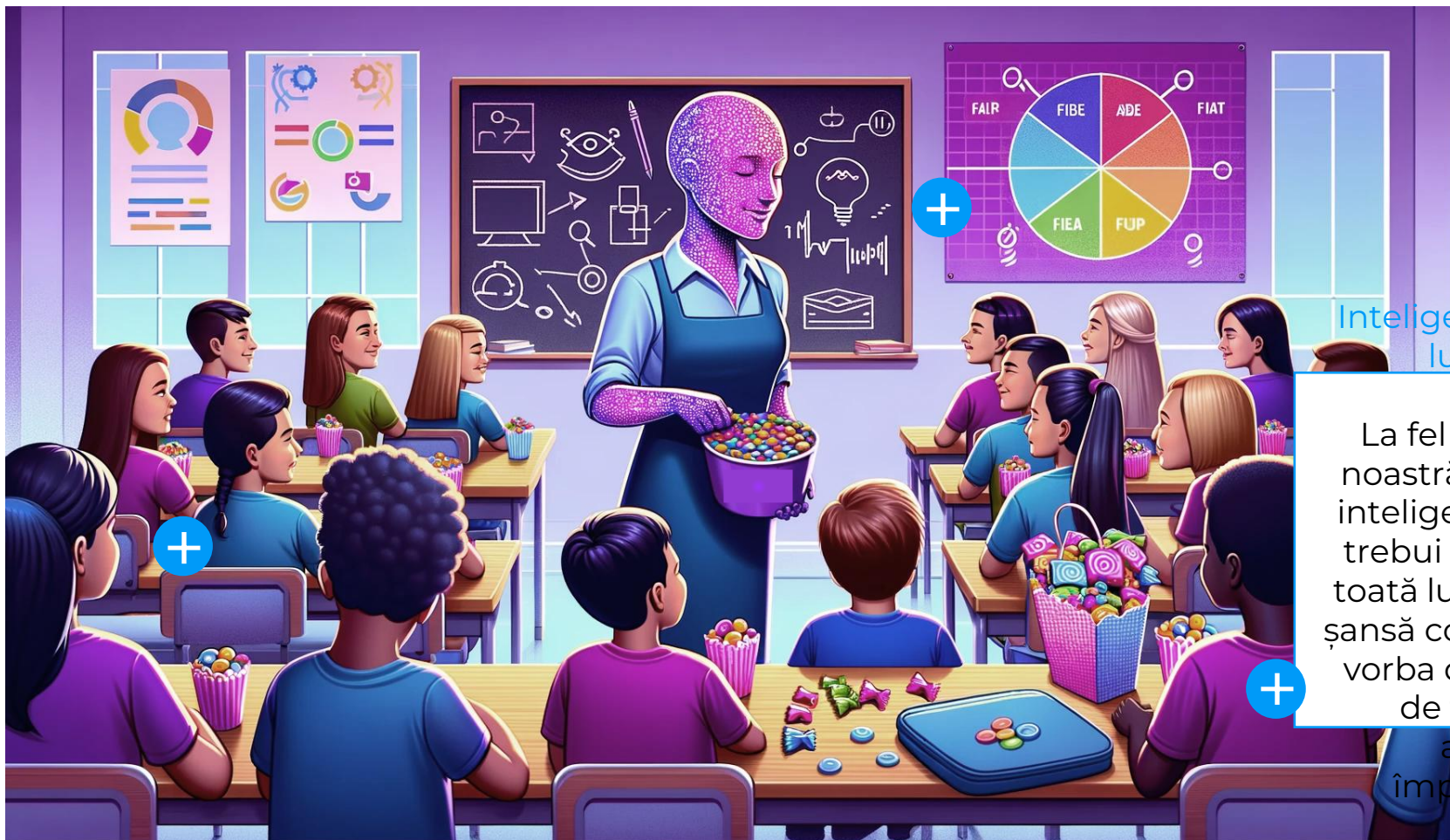
Introducere în corectitudinea și părtinirea datelor

Echitate și părtinire - analogia cu bomboanele



Introducere în corectitudinea și părtinirea datelor

Echitate și părtinire - analogia cu bomboanele



Inteligența artificială în lumea reală

La fel ca în povestea noastră cu bomboane, inteligența artificială ar trebui să se asigure că toată lumea primește o șansă corectă, fie că este vorba de recomandări de filme sau de aprobarea împrumuturilor.

Introducere în corectitudinea și părtinirea datelor

Corectitudinea datelor

Caracterul echitabil al datelor înseamnă asigurarea faptului că toate persoanele sunt tratate în mod echitabil de sistemele de inteligență artificială, fără ca prejudecățile să afecteze rezultatele acestora. Aceasta implică procesele și practicile care intervin în colectarea, prelucrarea și utilizarea datelor pentru a preveni efectele discriminatorii.

Corectitudinea datelor este un aspect crucial al proiectării și implementării sistemelor de inteligență artificială care au un impact pozitiv asupra societății. Aceasta garantează că deciziile AI sunt corecte și imparțiale, promovând egalitatea și prevenind discriminarea.

Practicile corecte de prelucrare a datelor conduc la aplicații AI demne de încredere, care pot fi acceptate pe scară largă și benefice pentru diverse comunități.

Scopul său este ca IA să deservească pe toată lumea în mod echitabil, în contrast cu trecutul care, din cauza caracteristicilor datelor, a fost

Introducere în corectitudinea și părtinirea datelor

Ce este părtinirea?	+
Cum se produce?	+
Care sunt consecințele prejudecăților?	+

Introducere în corectitudinea și părtinirea datelor

Ce este părtinirea?

Prejudecata în inteligența artificială se referă la erorile sistematice care favorizează în mod nedrept un grup în detrimentul altora. Aceasta poate rezulta din date distorsionate, algoritmi defectuoși sau procese de formare prejudiciate.

Cum se produce?

+

Care sunt consecințele prejudecăților?

+

Introducere în corectitudinea și părtinirea datelor

Ce este părtinirea?

+

Cum se produce?

-

Prejudicarea datelor: atunci când datele introduse în sistemele AI reflectă inegalități istorice sau reprezentări incomplete ale diverselor grupuri.

Prejudiciile algoritmului: atunci când regulile sau modelele utilizate pentru luarea deciziilor consolidează prejudiciile existente sau trec cu vederea tiparele minorităților.

Bucle de feedback: atunci când deciziile AI părtinitoare influențează datele viitoare, perpetuând ciclul părtinitor.

Care sunt consecințele prejudiciilor?

+

Introducere în corectitudinea și părtinirea datelor

Ce este părtinirea?

+

Cum se produce?

+

Care sunt consecințele prejudecăților?

-

Impactul social: poate duce la perpetuarea stereotipurilor, la consolidarea inegalităților sociale și la erodarea încrederii în instituțiile esențiale.

Impactul individual: oamenii se pot confrunta cu un tratament nedrept bazat pe decizii eronate ale AI în domenii critice precum angajarea, aplicarea legii și aprobarea împrumuturilor.

Introducere în corectitudinea și părtinirea datelor

Scenarii din lumea reală: Prejudecățile AI

Instrument de recrutare părtinitor

Scenariu

În 2018, s-a raportat că Amazon a eliminat un instrument de recrutare AI care arăta prejudecăți împotriva femeilor. Sistemul AI a fost antrenat pe baza datelor privind CV-urile colectate pe parcursul a 10 ani, predominant de la bărbați, reflectând dominația bărbaților în industria tehnologiei. În consecință, AI-ul a învățat singur că candidații bărbați sunt de preferat, penalizând CV-urile care includeau cuvinte precum "ale femeilor", așa cum ar putea apărea în "căpitanul clubului de șah al femeilor".

Impact

Acest exemplu evidențiază modul în care IA poate perpetua prejudecățile istorice în cazul în care datele de instruire sunt părtinitoare. De asemenea, arată potențialul IA de a afecta oportunitățile de angajare și importanța asigurării faptului că instrumentele de recrutare ale IA promovează diversitatea și echitatea.

Introducere în corectitudinea și părtinirea datelor

Scenarii din lumea reală: Prejudecățile AI

Prejudecarea algoritmului de asistență medicală

Scenariu

Un studiu din 2019 publicat de Science a arătat că un algoritm utilizat în multe sisteme de asistență medicală din SUA pentru alocarea resurselor de asistență medicală era părtinitor față de pacienții de culoare. Algoritmul a prezis nevoile de asistență medicală pe baza costurilor de asistență medicală, presupunând în mod inadvertit că costuri mai mici de asistență medicală înseamnă nevoi mai mici de asistență medicală. Cu toate acestea, din punct de vedere istoric, pacienții de culoare au înregistrat mai puține costuri din cauza diverselor bariere de acces la asistență medicală, nu pentru că erau mai sănătoși. Ca urmare, algoritmul a favorizat pacienții albi mai sănătoși în detrimentul pacienților negri mai bolnavi atunci când a

Impact

Această părtinire a algoritmului a făcut ca pacienții de culoare să beneficieze de un acces mult mai redus la programele de îngrijire concepute pentru a ajuta la gestionarea unor afecțiuni complexe, chiar dacă nevoia lor de astfel de intervenții era mai mare. Acest scenariu subliniază nevoia critică de echitate în aplicațiile IA din domeniul sănătății, unde luarea de decizii părtinitoare poate avea un impact direct asupra vieții și bunăstării.

TIPURI DE PREJUDECĂȚI ÎN SISTEMELE AI



Sursa imaginii | Generat de DALL-E

Tipuri de prejudecăți în sistemele AI

Selecția părtinitoare

Faceți clic pe cărți 
pentru a le
întoarce

Prejudecăți de
eșantionare

Faceți clic pe cărți 
pentru a le
întoarce

Tendința de măsurare

Faceți clic pe cărți 
pentru a le

Prejudecăți
algorithmice

Faceți clic pe cărți 
pentru a le
întoarce

Tendința de
confirmare

Faceți clic pe cărți 
pentru a le
întoarce

Tendința de raportare

Faceți clic pe cărți 
pentru a le
întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecăți de eșantionare

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendința de măsurare

Faceți clic pe cărți ↻
pentru a le
prejudecăți ale observatorului.

Prejudecăți algoritmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendința de confirmare

Faceți clic pe cărți ↻
pentru a le
dezvolta convingerile
întoarce

Tendința de raportare

Faceți clic pe cărți ↻
pentru a le
nefavorabile. întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecata de eșantionare este un subset al prejudecății de selecție și apare atunci când metoda utilizată pentru colectarea datelor favorizează anumite tipuri de persoane sau grupuri în detrimentul altora.

Tendința de măsurare

Faceți clic pe cărți ↻
pentru a le
prejudecăți ale observatorului.

Prejudecăți
algorithmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendința de
confirmare

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendința de raportare

Faceți clic pe cărți ↻
pentru a le
nefavorabile. întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecata de eșantionare este un subset al prejudecății de selecție și apare atunci când metoda utilizată pentru colectarea datelor favorizează anumite tipuri de persoane sau grupuri în detrimentul altora.

Prejudecățile de măsurare apar din cauza unor erori sau inconsecvențe în procesul de măsurare, ceea ce conduce la rezultate inexacte sau înșelătoare. Acestea pot proveni din instrumente defectuoase, întrebări ambigue sau prejudecăți ale observatorului.

Prejudecăți
algorithmice

Faceți clic pe cărți 
pentru a le
întoarce

Tendința de
confirmare

Faceți clic pe cărți 
pentru a le
întoarce

Tendința de raportare

Faceți clic pe cărți 
pentru a le
nefavorabile. întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecata de eșantionare este un subset al prejudecății de selecție și apare atunci când metoda utilizată pentru colectarea datelor favorizează anumite tipuri de persoane sau grupuri în detrimentul altora.

Prejudecățile de măsurare apar din cauza unor erori sau inconsecvențe în procesul de măsurare, ceea ce conduce la rezultate inexacte sau înșelătoare. Acestea pot proveni din instrumente defectuoase, întrebări ambigue sau prejudecăți ale observatorului.

Prejudecățile algoritmice apar în algoritmi de învățare automată atunci când aceștia discriminează în mod sistematic anumite grupuri sau persoane în funcție de rasă, sex sau alte caracteristici protejate prezente în datele de formare.

Tendința de confirmare

Faceți clic pe cărți
pentru a le
întoarce
întoarce

Tendința de raportare

Faceți clic pe cărți
pentru a le
nefavorabile. întoarce
întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecata de eșantionare este un subset al prejudecății de selecție și apare atunci când metoda utilizată pentru colectarea datelor favorizează anumite tipuri de persoane sau grupuri în detrimentul altora.

Prejudecățile de măsurare apar din cauza unor erori sau inconsecvențe în procesul de măsurare, ceea ce conduce la rezultate inexacte sau înșelătoare. Acestea pot proveni din instrumente defectuoase, întrebări ambigue sau prejudecăți ale observatorului.

Prejudecățile algoritmice apar în algoritmi de învățare automată atunci când aceștia discriminează în mod sistematic anumite grupuri sau persoane în funcție de rasă, sex sau alte caracteristici protejate prezente în datele de formare.

Tendința de confirmare atunci când cercetătorii sau analiștii utilizează sau interpretează în mod selectiv date care confirmă convingerile sau ipotezele lor preconcepute, ignorând sau respingând în același timp dovezile contradictorii.

Tendința de raportare

Faceți clic pe cărți pentru a le
nefavorabile. întoarce
întoarce

Tipuri de prejudecăți în sistemele AI

Abuzul de selecție apare atunci când eșantionul de date utilizat pentru analiză nu este reprezentativ pentru populația pe care ar trebui să o reprezinte. Aceasta conduce la concluzii distorsionate sau inexacte.

Prejudecata de eșantionare este un subset al prejudecății de selecție și apare atunci când metoda utilizată pentru colectarea datelor favorizează anumite tipuri de persoane sau grupuri în detrimentul altora.

Prejudecățile de măsurare apar din cauza unor erori sau inconsecvențe în procesul de măsurare, care conduc la rezultate inexacte sau înșelătoare. Acestea pot proveni din instrumente defectuoase, întrebări ambigue sau prejudecăți ale observatorului.

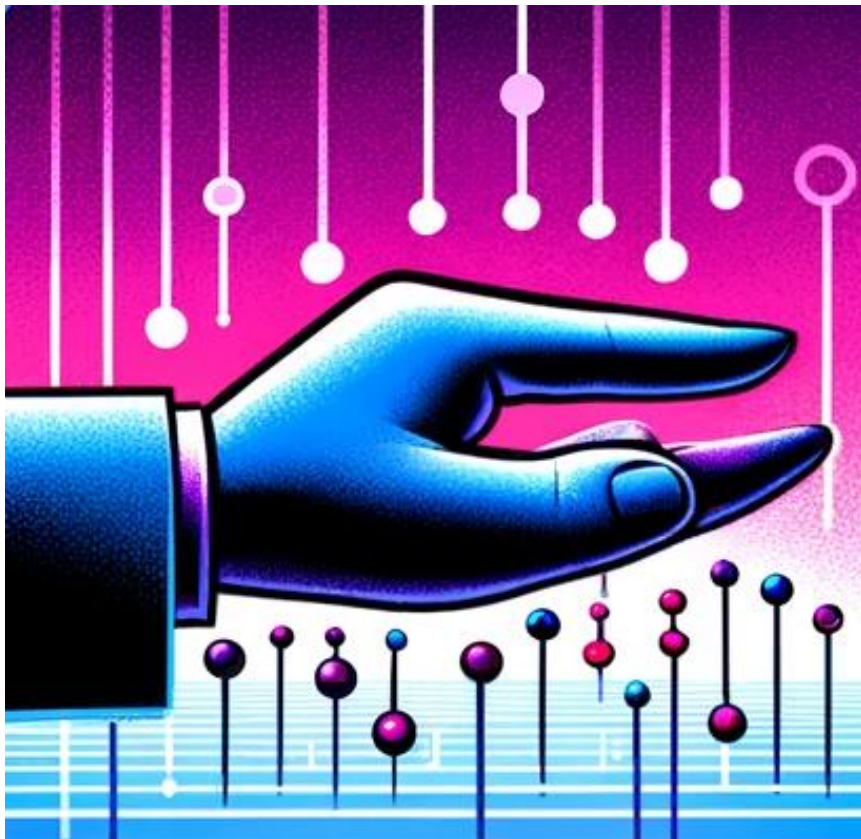
Prejudecățile algoritmice apar în algoritmi de învățare automată atunci când aceștia discriminează în mod sistematic anumite grupuri sau persoane în funcție de rasă, sex sau alte caracteristici protejate prezente în datele de formare.

Tendința de confirmare apare atunci când cercetătorii sau analiștii utilizează sau interpretează în mod selectiv date care confirmă convingerile sau ipotezele lor preconcepționale, ignorând sau respingând în același timp dovezile contradictorii.

Tendința de raportare se referă la publicarea sau raportarea selectivă a rezultatelor cercetării care sunt semnificative din punct de vedere statistic sau favorabile, suprimând sau neglijând rezultatele nesemnificative sau nefavorabile.

Tipuri de prejudecăți în sistemele AI

Exemple



Sursa imaginii | Generat de DALL-E

Selecția părtinitoare

Tendințe de selecție | Tehnologia de recunoaștere facială

Tehnologiile de recunoaștere facială au fost adesea antrenate predominant pe seturi de date care cuprind mai ales fețe cauziene. Această lipsă de diversitate a seturilor de date de instruire a condus la rate de eroare mai ridicate în recunoașterea persoanelor din alte medii etnice. De exemplu, un studiu realizat de MIT Media Lab a constatat disparități semnificative în ceea ce privește acuratețea tehnologiilor de recunoaștere a sexului, în special identificarea greșită a sexului în rândul femeilor cu pielea mai închisă la culoare.

Tipuri de prejudecăți în sistemele AI

Exemple



Prejudecăți de eșantionare | Erorile sondajelor politice

În cadrul alegerilor prezidențiale din 2016 din SUA, mai multe sondaje de opinie au subestimat sprijinul pentru Donald Trump. O problemă cheie a fost prejudecata de eșantionare, în care ponderea demografică în eșantioane nu a reprezentat cu exactitate baza electorală. Sondajele au supraeșantionat populațiile urbane și au subeșantionat zonele rurale, ceea ce a condus la predicții inexacte care nu au ținut rezultatul electoral real.

Sursa imaginii | Generat de DALL-E

Tipuri de prejudecăți în sistemele AI

Exemple



Tendința de măsurare | Pulsoximetre și tonul pielii

Studii recente, inclusiv cele intensificate de pandemia COVID-19, au arătat că oximetrele de puls pot prezenta erori de măsurare prin supraestimarea nivelului de oxigen din sânge la persoanele cu pielea mai închisă la culoare. Această tendință apare deoarece senzorii dispozitivelor sunt mai puțin eficienți la penetrarea pielii mai închise la culoare, ceea ce duce la diagnostice eronate potențial periculoase sau la întârzieri în tratament pentru pacienții de culoare.

Sursa imaginii | Generat de DALL-E Tendința de măsurare

Tipuri de prejudecăți în sistemele AI

Exemple



Sursa imaginii | Generat de DALL-E
Prejudecăți algoritmice

Prejudecăți algoritmice | COMPAS în condamnarea penală

S-a constatat că software-ul COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizat în sistemul de justiție penală din SUA pentru a evalua probabilitatea ca un inculpat să recidiveze, prezintă prejudecăți rasiale. Analiza a arătat că software-ul avea o probabilitate de aproape două ori mai mare de a prezice în mod eronat criminalitatea viitoare a inculpaților de culoare în comparație cu inculpații albi, influențând deciziile de condamnare și de eliberare pe cauțiune pe baza unor recomandări algoritmice distorsionate.

Tipuri de prejudecăți în sistemele AI

Exemple



Sursa imaginii | Generat de DALL-E Tendința de confirmare

Biasa de confirmare - Modele de scoring al creditelor

Instituțiile financiare care utilizează modele de scoring al creditelor bazate pe inteligența artificială ar putea consolida în mod neintenționat prejudecățile de confirmare dacă modelele sunt antrenate pe baza datelor istorice de creditare care reflectă prejudecățile din trecut. De exemplu, dacă creditorii umani anteriori au avut prejudecăți față de anumite grupuri demografice, sistemul AI ar putea reproduce aceste prejudecăți favorizând persoanele din grupurile favorizate istoric, presupunând că acestea sunt mai puțin riscante pe baza rambursărilor anterioare ale împrumuturilor.

Tipuri de prejudecăți în sistemele AI

Exemple



Sursa imaginii | Generat de DALL-E Tendința de raportare

Raportarea părtinitoare - Efectele secundare ale medicamentelor în forumurile online

Companiile farmaceutice și cercetătorii din domeniul sănătății care utilizează inteligența artificială pentru a monitoriza forumurile online și rețelele de socializare pentru rapoartele privind efectele secundare ale medicamentelor se pot confrunta cu o părtinire a raportării. Pacienții care se confruntă cu efecte secundare grave sunt mai predispuși să își raporteze experiențele online în comparație cu cei care se confruntă cu efecte secundare ușoare sau fără efecte secundare. Acest lucru poate determina sistemele AI să supraestimeze gravitatea sau frecvența reacțiilor adverse, putând influența în mod inexact profilurile de siguranță ale medicamentelor.

METODE DE IDENTIFICARE ȘI MĂSURARE A PĂRTINIRILOR ÎN DATE



Sursa imaginii | Generat de DALL-E

Metode de identificare și măsurare a părtinirilor din date

Auditul datelor	+
Analiza disparităților	+
Testarea algoritmică a corectitudinii	+
Analiza sensibilității	+

Metode de identificare și măsurare a părtinirilor din date

Auditul datelor

-

Auditul datelor implică examinarea sistematică a seturilor de date pentru a identifica orice disparități, dezechilibre sau anomalii care pot indica prejudecăți.

Acest proces include verificarea egalității de reprezentare între diferitele grupuri din cadrul datelor. Se asigură că procesul de colectare a datelor nu exclude sistematic anumite populații sau nu consolidează stereotipurile existente.

Analiza disparităților

+

Testarea algoritmică a corectitudinii

+

Analiza de sensibilitate

+

Metode de identificare și măsurare a părtinirilor din date

Auditul datelor	+
Analiza disparităților Analiza disparităților se referă la compararea rezultatelor sau a predicțiilor modelelor de inteligență artificială între diferite grupuri demografice sau protejate pentru a identifica orice impact inegal. Aceasta implică calcularea unor măsuri statistice, cum ar fi diferența dintre ratele de eroare, ratele de acceptare sau alți indicatori de rezultat între grupuri. Acest lucru ajută la cuantificarea și vizualizarea disparităților care pot indica prejudecăți în procesele decizionale ale IA.	-
Testarea algoritmică a corectitudinii	+
Analiza sensibilității	+

Metode de identificare și măsurare a părtinirilor din date

Auditul datelor	+
Analiza disparităților	+
Testarea algoritmică a corectitudinii Testarea corectitudinii algoritmice implică punerea în aplicare a testelor sau a parametrilor care evaluează corectitudinea deciziilor luate de sistemele AI. Aceasta poate include utilizarea unor parametri de echitate, cum ar fi egalitatea de șanse, paritatea demografică sau criteriile de echitate individuală pentru a evalua performanța algoritmilor. Aceste teste ajută la determinarea faptului dacă sistemele AI iau decizii corecte și imparțiale în cadrul diferitelor grupuri demografice.	-
Analiza de sensibilitate	+

Metode de identificare și măsurare a părtinirilor din date

Auditul datelor	+
Analiza disparităților	+
Testarea algoritmică a corectitudinii	+
Analiza sensibilității Analiza de sensibilitate presupune testarea solidității deciziilor AI prin modificarea ușoară a datelor de intrare și observarea modului în care rezultatele variază pentru diferite grupuri. Prin perturbarea datelor de intrare, analiza sensibilității evaluează cât de sensibile sunt modelele AI la modificările datelor de intrare. Aceasta ajută la identificarea surselor potențiale de erori și evaluează fiabilitatea și coerența predicțiilor AI în diverse seturi de date.	-

Metode de identificare și măsurare a părtinirilor din date

Atelier interactiv pentru identificarea și măsurarea prejudecăților din date Instrument de recrutare părtinitor

Obiectiv: acest atelier va permite participanților să aplice auditul datelor, analiza disparităților și analiza sensibilității utilizând un set de date reale pentru a identifica și înțelege potențialele prejudecăți.

Unelte și materiale

Set de date: furnizați un set de date simplu, disponibil publicului, cu diverse atribute demografice (de exemplu, setul de date Adult Income din UCI Machine Learning Repository, care include caracteristici precum vârsta, educația, rasa, sexul și venitul).

Activitate: participanții încarcă setul de date și efectuează explorarea de bază a datelor (verificarea valorilor lipsă și înțelegerea distribuției atributelor cheie).

Scopul: familiarizarea participanților cu setul de date și pregătirea datelor pentru analiză.

Auditul datelor: efectuați un audit al datelor prin analizarea reprezentării diferitelor grupuri în cadrul setului de date. Calculați proporția fiecărui grup demografic (de exemplu, în funcție de sex, rasă). Identificați orice grupuri subreprezentate.

Analiza disparităților: participanții aplică analiza disparităților pentru a evalua măsura în care un model simplu (de exemplu, regresia logistică pentru prezicerea nivelului de venit) prezice diferite grupuri. Identifică orice disparități semnificative în performanța modelului între grupuri și formulează ipoteze privind cauzele potențiale.

Analiza sensibilității: efectuați o analiză a sensibilității prin modificarea ușoară a punctelor de date și observarea schimbărilor în predicțiile modelului. Modificați atributele legate de variabilele demografice sensibile (de exemplu, modificați ușor vârsta sau nivelul de educație) și observați modul în care modificările afectează predicțiile modelului.

STRATEGII PENTRU ABORDAREA ȘI ATENUAREA PREJUDECĂȚILOR



Sursa imaginii | Generat de DALL-E

Strategii pentru abordarea și atenuarea prejudecăților

Reprezentare diversă

Faceți clic pe cărți ↻
pentru a le
întoarce

Monitorizare și evaluare
periodică

Faceți clic pe cărți ↻
pentru a le
întoarce

Algoritmi care țin seama de
prejudecăți

Faceți clic pe cărți ↻
pentru a le
întoarce

Transparență și explicabilitate

Faceți clic pe cărți ↻
pentru a le
întoarce

Strategii pentru abordarea și atenuarea prejudecăților

Reprezentare diversă

Diversitatea punctelor de date din diferite grupuri demografice asigură faptul că sistemele AI sunt antrenate pe un set de date reprezentativ, reducând probabilitatea unor rezultate părtinitoare. Diversitatea aduce perspective și idei diferite, ajutând la identificarea și atenuarea prejudecăților în timpul procesului de dezvoltare.

Monitorizare și evaluare periodică

Faceți clic pe cărți
pentru a le
întoarce

Algoritmi care țin seama de prejudecăți

Faceți clic pe cărți
pentru a le
întoarce

Transparență și explicabilitate

Faceți clic pe cărți
pentru a le
întoarce

rezultatelor inteligenței artificiale.

Strategii pentru abordarea și atenuarea prejudecăților

Reprezentare diversă

Diversitatea punctelor de date din diferite grupuri demografice asigură faptul că sistemele AI sunt antrenate pe un set de date reprezentativ, reducând probabilitatea unor rezultate părtinitoare. Diversitatea aduce perspective și idei diferite, ajutând la identificarea și atenuarea prejudecăților în timpul procesului de dezvoltare.

Monitorizare și evaluare
periodică

Faceți clic pe cărți
pentru a le
întoarce

Algoritmi care țin seama de prejudecăți

Conceperea de algoritmi cu mecanisme integrate pentru detectarea și atenuarea prejudecăților. Acest lucru poate implica încorporarea constrângerilor de corectitudine în procesul de optimizare a algoritmului sau utilizarea unor tehnici precum instruirea adversarială pentru a minimiza prejudecățile în predicțiile

Transparență și explicabilitate

Faceți clic pe cărți
pentru a le
rezultatelor inteligenței artificiale.
întoarce

Strategii pentru abordarea și atenuarea prejudecăților

Reprezentare diversă

Diversitatea punctelor de date din diferite grupuri demografice asigură faptul că sistemele AI sunt antrenate pe un set de date reprezentativ, reducând probabilitatea unor rezultate părtinitoare. Diversitatea aduce perspective și idei diferite, ajutând la identificarea și atenuarea prejudecăților în timpul procesului de dezvoltare.

Monitorizare și evaluare periodică

Monitorizarea continuă a sistemelor AI în contexte reale pentru a identifica prejudecățile pe măsură ce acestea apar. Evaluarea periodică implică analiza rezultatelor AI în ceea ce privește disparitățile între diferite grupuri demografice și actualizarea algoritmilor sau a seturilor de date pentru a aborda prompt orice prejudecăți identificate.

Algoritmi care țin seama de prejudecăți

Conceperea de algoritmi cu mecanisme integrate pentru detectarea și atenuarea prejudecăților. Acest lucru poate implica încorporarea constrângerilor de corectitudine în procesul de optimizare a algoritmului sau utilizarea unor tehnici precum instruirea adversarială pentru a minimiza prejudecățile în predicțiile

Transparență și explicabilitate

Faceți clic pe cărți

pentru a le
rezultatelor inteligenței artificiale.
întoarce

Strategii pentru abordarea și atenuarea prejudecăților

Reprezentare diversă

Diversitatea punctelor de date din diferite grupuri demografice asigură faptul că sistemele AI sunt antrenate pe un set de date reprezentativ, reducând probabilitatea unor rezultate părtinitoare. Diversitatea aduce perspective și idei diferite, ajutând la identificarea și atenuarea prejudecăților în timpul procesului de dezvoltare.

Monitorizare și evaluare periodică

Monitorizarea continuă a sistemelor AI în contexte reale pentru a identifica prejudecățile pe măsură ce acestea apar. Evaluarea periodică implică analiza rezultatelor AI în ceea ce privește disparitățile între diferite grupuri demografice și actualizarea algoritmilor sau a seturilor de date pentru a aborda prompt orice prejudecăți identificate.

Algoritmi care țin seama de prejudecăți

Conceperea de algoritmi cu mecanisme integrate pentru detectarea și atenuarea prejudecăților. Acest lucru poate implica încorporarea constrângerilor de corectitudine în procesul de optimizare a algoritmului sau utilizarea unor tehnici precum instruirea adversarială pentru a minimiza prejudecățile în predicțiile

modelului. Transparență și explicabilitate

Oferirea de transparență și explicații pentru deciziile AI permite părților interesate să înțeleagă modul în care sunt luate deciziile și să identifice orice prejudecăți prezente în sistem. Tehnicile de inteligență artificială explicabile, cum ar fi furnizarea de scoruri de importanță a caracteristicilor sau de justificări ale deciziilor, permit utilizatorilor să interpreteze și să evalueze corectitudinea rezultatelor inteligenței artificiale.

ÎNȚELEGEREA PARAMETRILOR DE CORECTITUDINE



Sursa imaginii | Generat de DALL-E

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale	+
Disparitatea demografică condiționată	+
Echitate prin conștientizare	+
Echitate individuală	+
Echitate contrafactuală	+

Înțelegerea parametrilor de corectitudine

Paritatea statistică

Paritatea statistică, cunoscută și sub denumirea de paritate demografică, evaluează dacă proporția de rezultate pozitive (de exemplu, aprobări de credite sau oferte de locuri de muncă) este aceeași în cadrul diferitelor grupuri demografice.

Șanse egale

+

Disparitatea demografică condiționată

+

Echitate prin conștientizare

+

Echitate individuală

+

Echitate contrafactuală

+

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale Cotele egalizate examinează dacă rata pozitivă reală (sensibilitate) și rata negativă reală (specificitate) sunt egale în cadrul diferitelor grupuri demografice. Aceasta garantează că performanța predictivă a modelului este consecventă în toate grupurile.	
Disparitatea demografică condiționată	+
Echitate prin conștientizare	+
Echitate individuală	+
Echitate contrafactuală	+

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale	+
Disparitatea demografică condiționată: Disparitatea demografică condiționată evaluează disparitatea în rezultatele predictive condiționate de un atribut protejat. Aceasta măsoară diferența în rezultatele predicțiilor între diferite grupuri demografice, ținând seama de alți factori relevanți.	
Echitate prin conștientizare	+
Echitate individuală	+
Echitate contrafactuală	+

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale	+
Disparitatea demografică condiționată	+
Echitate prin conștientizare Echitatea prin conștientizare încorporează conștientizarea atributelor sensibile (de exemplu, rasa sau sexul) în procesul decizional. Aceasta garantează că deciziile luate de sistemele AI sunt corecte și imparțiale, luând în considerare contextul istoric și impactul asupra societății.	
Echitate individuală	+
Echitate contrafactuală	+

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale	+
Disparitatea demografică condiționată	+
Echitate prin conștientizare	+
Echitate individuală Echitatea individuală se concentrează pe asigurarea faptului că persoane similare primesc un tratament sau rezultate similare din partea sistemului IA, indiferent de caracteristicile lor demografice. Scopul este de a minimiza tratamentul diferențiat bazat pe atribute irelevante.	
Echitate contrafactuală	+

Înțelegerea parametrilor de corectitudine

Paritatea statistică	+
Șanse egale	+
Disparitatea demografică condiționată	+
Echitate prin conștientizare	+
Echitate individuală	+
Echitate contrafactuală Echitatea contrafactuală evaluează dacă modificarea unui atribut sensibil al unei persoane (de exemplu, rasa sau sexul), păstrând constante celelalte caracteristici, ar duce la o modificare a rezultatului deciziei. Aceasta garantează că persoanele sunt tratate în mod echitabil, indiferent de atributele lor protejate.	

Înțelegerea parametrilor de corectitudine

Test interactiv

Crearea unui test interactiv pe Mentimeter

Întrebare sondaj

Înainte de a începe, care sunt factorii importanți care trebuie luați în considerare pentru a determina dacă un sistem de inteligență artificială este corect?

Opțiuni

a) echitatea rezultatelor; b) reprezentarea demografică; c) tratamentul egal; d) transparența.

Înțelegerea parametrilor de corectitudine

Exerciții interactive în clasă

Subiect: Paritatea statistică

Calcularea parității statistice

- Date fiind datele, calculați rata de aprobare a împrumuturilor pentru fiecare rasă. Există paritate?
- Furnizați participanților numerele (sau procente) de aprobări și respingeri pentru fiecare rasă și cereți-le să calculeze dacă ratele sunt aproximativ egale.
- Răspunsuri cu opțiuni multiple: "Ratele sunt egale; paritatea este atinsă." și "Ratele nu sunt egale; există o disparitate."

Înțelegerea parametrilor de corectitudine

Exerciții interactive în clasă

Subiect: Șanse egale

Evaluarea șanselor egalizate

- Presupunând că rata pozitivă reală (împrumut corect aprobat pentru cei care ar trebui să fie aprobați) pentru albi este de 80%, iar pentru negri este de 60%, există șanse egale?
- Răspunsuri cu opțiuni multiple: "Da, șansele sunt egale." sau "Nu, șansele nu sunt egale."

CORECTITUDINE ÎN COLECTAREA ȘI PREPROCESAREA DATELOR



Sursa imaginii | Generat de DALL-E

Corectitudine în colectarea și preprocesarea datelor



Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Faceți clic pe cărți

minoritate pentru a le
e. întoarce

Curățarea datelor

Faceți clic pe cărți

pentru a le
întoarce

Detectarea deviațiilor

Faceți clic pe cărți

disparitățile între diferite grupuri
întoarce

Ingineria
caracteristicilor
echitabile

Faceți clic pe cărți

pentru a le
folosiți caracteristici care sunt
relevante și nediscriminatorii.

Transparență și
documentare

Faceți clic pe cărți

pentru a le
avea un impact asupra corectitudinii.
întoarce

Monitorizare
periodică

Faceți clic pe cărți

pentru a le
preprocesare pentru a asigura
echitatea în timp.
întoarce

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Faceți clic pe cărți
pentru a le
întoarce

Detectarea deviațiilor

Faceți clic pe cărți
pentru a le
disparitățile între diferite grupuri
întoarce

Ingineria caracteristicilor echitabile

Faceți clic pe cărți
pentru a le
folosiți caracteristici care sunt
relevante și nediscriminatorii.

Transparență și documentare

Faceți clic pe cărți
pentru a le
avea un impact asupra corectitudinii.
întoarce

Monitorizare periodică

Faceți clic pe cărți
pentru a le
preprocesare pentru a asigura
echitatea în timp.
întoarce

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Curățați și preprocesați cu atenție datele pentru a elimina orice prejudecăți sau neconcordanțe. Aceasta include identificarea și corectarea erorilor, gestionarea valorilor lipsă și standardizarea formatelor de date.

Detectarea deviațiilor

Faceți clic pe cărți

pentru a le
disparitățile între diferite grupuri
întoarce

Ingineria caracteristicilor echitabile

Faceți clic pe cărți

pentru a le
folosiți caracteristici care sunt
relevante și nediscriminatorii.

Transparență și documentare

Faceți clic pe cărți

pentru a le
avea un impact asupra corectitudinii.
întoarce

Monitorizare periodică

Faceți clic pe cărți

pentru a le
preprocesare pentru a asigura
echitatea în timp.
întoarce

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Curățați și preprocesați cu atenție datele pentru a elimina orice prejudecăți sau neconcordanțe. Aceasta include identificarea și corectarea erorilor, gestionarea valorilor lipsă și standardizarea formatelor de date.

Detectarea deviațiilor

Implementați tehnici de detectare și atenuare a prejudecăților din date. Aceasta poate implica efectuarea de audituri ale prejudecăților sau utilizarea de instrumente automatizate pentru a identifica disparitățile între diferite grupuri

Ingineria
caracteristicilor
echitabile

Faceți clic pe cărți pentru a le

folosiți caracteristici care sunt relevante și nediscriminatorii.

Transparență și
documentare

Faceți clic pe cărți pentru a le

avea un impact asupra corectitudinii.

Monitorizare
periodică

Faceți clic pe cărți pentru a le

preprocesa pentru a asigura echitatea în timp.

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Curățați și preprocesați cu atenție datele pentru a elimina orice prejudecăți sau neconcordanțe. Aceasta include identificarea și corectarea erorilor, gestionarea valorilor lipsă și standardizarea formatelor de date.

Detectarea deviațiilor

Implementați tehnici de detectare și atenuare a prejudecăților din date. Aceasta poate implica efectuarea de audituri ale prejudecăților sau utilizarea de instrumente automatizate pentru a identifica disparitățile între diferite grupuri

Ingineria caracteristicilor echitabile

Luați în considerare implicațiile selecției și ingineriei caracteristicilor asupra corectitudinii. Evitați utilizarea caracteristicilor care pot perpetua stereotipurile sau discriminarea și străduiți-vă să folosiți caracteristici care sunt relevante și nediscriminatorii.

Transparență și documentare

Faceți clic pe cărți pentru a le avea un impact asupra corectitudinii. întoarce

Monitorizare periodică

Faceți clic pe cărți pentru a le preprocesa pentru a asigura echitatea în timp.

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Curățați și preprocesați cu atenție datele pentru a elimina orice prejudecăți sau neconcordanțe. Aceasta include identificarea și corectarea erorilor, gestionarea valorilor lipsă și standardizarea formatelor de date.

Detectarea deviațiilor

Implementați tehnici de detectare și atenuare a prejudecăților din date. Aceasta poate implica efectuarea de audituri ale prejudecăților sau utilizarea de instrumente automatizate pentru a identifica disparitățile între diferite grupuri

Ingineria caracteristicilor echitabile

Luați în considerare implicațiile selecției și ingineriei caracteristicilor asupra corectitudinii. Evitați utilizarea caracteristicilor care pot perpetua stereotipurile sau discriminarea și străduiți-vă să folosiți caracteristici care sunt relevante și nediscriminatorii.

Transparență și documentare

Documentați procedurile de colectare și preprocesare a datelor pentru a asigura transparența și responsabilitatea. Aceasta include documentarea oricăror decizii luate în timpul curățării datelor sau al ingineriei caracteristice care ar putea avea un impact asupra corectitudinii.

Monitorizare periodică

Faceti clic pe căști pentru a le preprocesa pentru a asigura echitatea în timp.

Corectitudine în colectarea și preprocesarea datelor

Diverse surse de date

Colectați date din diverse surse pentru a asigura o reprezentare cuprinzătoare a populației. Acest lucru ajută la prevenirea sub-reprezentării sau a prejudecăților față de anumite grupuri demografice.

Eșantionare incluzivă

Utilizați tehnici de eșantionare incluzive pentru a vă asigura că toate segmentele populației sunt reprezentate în mod adecvat în setul de date. Aceasta poate implica eșantionarea stratificată sau supraeșantionarea grupurilor minoritare pentru a remedia

Curățarea datelor

Curățați și preprocesați cu atenție datele pentru a elimina orice prejudecăți sau neconcordanțe. Aceasta include identificarea și corectarea erorilor, gestionarea valorilor lipsă și standardizarea formatelor de date.

Detectarea deviațiilor

Implementați tehnici de detectare și atenuare a prejudecăților din date. Acest lucru poate implica efectuarea de audituri privind prejudecățile sau utilizarea de instrumente automatizate pentru a identifica disparitățile între diferite grupuri

Ingineria caracteristicilor echitabile

Luați în considerare implicațiile selecției și ingineriei caracteristicilor asupra corectitudinii. Evitați utilizarea caracteristicilor care pot perpetua stereotipurile sau discriminarea și străduiți-vă să folosiți caracteristici care sunt relevante și nediscriminatorii.

Transparență și documentare

Documentați procedurile de colectare și preprocesare a datelor pentru a asigura transparența și responsabilitatea. Aceasta include documentarea oricăror decizii luate în timpul curățării datelor sau al ingineriei caracteristice care ar putea avea un impact asupra corectitudinii.

Monitorizare periodică

Monitorizarea continuă a procesului de colectare și preprocesare a datelor pentru a identifica și aborda orice prejudecăți sau disparități emergente. Actualizați periodic setul de date și tehnicile de preprocesare pentru a asigura echitatea în timp.

TENDINȚE VIITOARE ȘI PROBLEME EMERGENTE



Sursa imaginii | Generat de DALL-E

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Faceți clic pe cărți ↻
pentru a le
întoarce

Atacuri și apărări
adversare

Faceți clic pe cărți ↻
pentru a le
întoarce

Echitate în mai multe
dimensiuni

Faceți clic pe cărți ↻
pentru a le
întoarce

Atenuarea
prejudecăților
algorithmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Considerații etice și
guvernanță

Faceți clic pe cărți ↻
pentru a le
întoarce

Proiectare centrată
pe om

Faceți clic pe cărți ↻
pentru a le
întoarce

Echitate în
tehnologiile
emergente

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și interpretabilitate

Pe măsură ce sistemele AI devin mai complexe, există o cerere tot mai mare de transparență și interpretabilitate. Tendințele viitoare s-ar putea concentra pe dezvoltarea de tehnici și standarde pentru explicarea deciziilor AI și pentru a le face mai ușor de interpretat pentru părțile interesate.

Atacuri și apărări
adversare

Faceți clic pe cărți ↻
pentru a le
întoarce

Echitate în mai multe
dimensiuni

Faceți clic pe cărți ↻
pentru a le
întoarce

Atenuarea
prejudecăților
algorithmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Considerații etice și
guvernanță

Faceți clic pe cărți ↻
pentru a le
întoarce

Proiectare centrată
pe om

Faceți clic pe cărți ↻
pentru a le
întoarce

Echitate în
tehnologiile
emergente

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări adversare

Atacurile adversariale, în care actorii rău intenționați manipulează sistemele AI prin modificarea subtilă a datelor de intrare, reprezintă o amenințare semnificativă la adresa corectitudinii și robusteții. Cercetările viitoare ar putea explora mijloace de apărare și contramăsuri robuste împotriva atacurilor adversariale pentru a spori

Echitate în mai multe
dimensiuni

Faceți clic pe cărți ↻
pentru a le

Atenuarea
prejudecăților
algorithmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Considerații etice și
guvernare

Faceți clic pe cărți ↻
pentru a le

întoarce

Proiectare centrată
pe om

Faceți clic pe cărți ↻
pentru a le

întoarce

Echitate în
tehnologiile
emergente

Faceți clic pe cărți ↻
pentru a le

întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări
adversare

Echitate în mai multe dimensiuni

Echitatea în inteligența artificială merge dincolo de atributele demografice și poate cuprinde alte dimensiuni, cum ar fi statutul socioeconomic, dizabilitatea sau competențele lingvistice. Tendințele viitoare ar putea implica dezvoltarea de măsurători și algoritmi de corectitudine multidimensională pentru a aborda prejudecățile intersecționale.

Atenuarea
prejudecăților
algorithmice

Faceți clic pe cărți ↻
pentru a le
întoarce

Considerații etice și
guvernanță

Faceți clic pe cărți ↻
pentru a le
întoarce

Proiectare centrată
pe om

Faceți clic pe cărți ↻
pentru a le
întoarce

Echitate în
tehnologiile
emergente

Faceți clic pe cărți ↻
pentru a le
întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări
adversare

Echitate în mai multe
dimensiuni

Atenuarea prejudecăților algoritmice

Progresele în tehnicile algoritmice de atenuare a prejudecăților vor continua să fie un punct central, cercetarea concentrându-se pe dezvoltarea unor algoritmi și modele mai corecte care reduc prejudecățile în diverse contexte decizionale.

Considerații etice și
guvernanță

Proiectare centrată
pe om

Echitate în
tehnologiile
emergente

Faceți clic pe cărți ↻
pentru a le

întoarce

Faceți clic pe cărți ↻
pentru a le

întoarce

Faceți clic pe cărți ↻
pentru a le

întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări
adversare

Echitate în mai multe
dimensiuni

Atenuarea
prejudecăților
algorithmice

Considerații etice și guvernare

Implicațiile etice ale corectitudinii și prejudecăților IA vor rămâne un domeniu cheie de preocupare, cu un accent tot mai mare pe practicile etice de dezvoltare a IA și pe cadrele de guvernare. Tendințele viitoare ar putea implica stabilirea de orientări și standarde de reglementare pentru asigurarea corectitudinii și responsabilității în sistemele de inteligență artificială.

Proiectare centrată
pe om

Faceți clic pe cărți
pentru a le
întoarce

Echitate în
tehnologiile
emergente

Faceți clic pe cărți
pentru a le
întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări
adversare

Echitate în mai multe
dimensiuni

Atenuarea
prejudecăților
algorithmice

Considerații etice și
guvernanță

Proiectare centrată pe om

Viitoarele sisteme de inteligență artificială pot acorda prioritate principiilor de proiectare centrate pe om, încorporând feedback-ul și preferințele utilizatorilor pentru a spori corectitudinea și utilitatea. Acest lucru ar putea implica coproiectarea sistemelor AI cu diverse părți interesate pentru a asigura incluziunea și echitatea.

Echitate în
tehnologiile
emergente

Faceți clic pe cărți
pentru a le
întoarce

Tendințe viitoare și probleme emergente

Explicabilitate și
interpretabilitate

Atacuri și apărări
adversare

Echitate în mai multe
dimensiuni

Atenuarea
prejudecăților
algorithmice

Considerații etice și
guvernanță

Proiectare centrată
pe om

Echitate în tehnologiile emergente

Pe măsură ce inteligența artificială continuă să se intersecteze cu tehnologii emergente, cum ar fi vehiculele autonome, inteligența artificială în domeniul sănătății și sistemele de recunoaștere facială, abordarea echității și a prejudecăților devine din ce în ce mai importantă. Tendințele viitoare pot implica adaptarea principiilor echității la noile aplicații și progrese tehnologice.

CONCLUZII



Sursa imaginii | Generat de DALL-E

CONCLUZII

1. **Înțelegerea complexității:** Egalitatea și prejudecățile în materie de IA sunt provocări cu multiple fațete care necesită strategii cuprinzătoare și o atenție permanentă.
2. **Strategii de acțiune:** de la reprezentarea diversă în colectarea datelor la atenuarea prejudecăților algoritmice, este disponibilă o serie de strategii pentru abordarea și atenuarea prejudecăților în sistemele AI.
3. **Anticiparea provocărilor emergente:** pe măsură ce inteligența artificială continuă să evolueze, va fi esențial să rămânem înaintea tendințelor și problemelor emergente, cum ar fi atacurile adversarilor, corectitudinea multidimensională și considerentele etice.
4. **Către un viitor mai echitabil:** prin prioritizarea designului centrat pe om, a transparenței și a responsabilității, putem lucra la construirea unor sisteme AI care să promoveze corectitudinea, incluziunea și încrederea.

VĂ MULTUMESC



Universitat
de les Illes Balears

ISQe
ENGAGING PEOPLE



AARHUS
UNIVERSITY

VAMK
VAASAN AMMATTIKORKEAKOULU
UNIVERSITY OF APPLIED SCIENCES

helixconnect



Cofinanțat de
Uniunea Europeană

Finanțat de Uniunea Europeană. Punctele de vedere și opiniile exprimate aparțin, însă, exclusiv autorului (autorilor) și nu reflectă neapărat punctele de vedere și opiniile Uniunii Europene sau ale Agenției Executive Europene pentru Educație și Cultură (EACEA). Nici Uniunea Europeană și nici EACEA nu pot fi considerate răspunzătoare pentru acestea.

CharlTe

2022-1-ES01-KA220-HED-000085257