Fișă de învățare

Titlu	Data Science & Impactul Social: Obținerea rezultatelor pozitive
Cuvinte cheie(meta tag)	Impact Social, Date pentru binele general, metrici de echitate, monitorizarea social media
Limbă	Română
Obiective / Scop / Rezultatele învățării	**1. Utilizarea data science pentru binele social** **2. Înțelegerea principalelor riscuri ale tehnologiei și identificarea exemplelor** **3. Să fiți capabili să enumerați caracteristicile ”AI de încredere”** **4. Să înțelegeți provocările măsurării echității**

Curs:	
Data Science Literacy	
Vizualizarea Datelor și Modulul de Visual Analytics	
Introducere în Data science pentru Științe sociale	
Data Science pentru binele public	X
Jurnalismul de date și Storytelling	

Descriere	În acest curs, vom arunca o privire asupra numeroaselor aplicații ale Data Science care pot face lumea un loc mai bun. Vom intra apoi în detaliu asupra monitorizării rețelelor sociale efectuate în numele Amnesty International Italia, pentru a înțelege cum poate funcționa o astfel de aplicație. În secțiunea următoare, vom explora unele dintre efectele dăunătoare pe care le pot avea Data Science și AI (Inteligența Artificială). Acest lucru ne va ajuta să înțelegem de ce este nevoie ca sistemele AI să fie de încredere. În cele din urmă, ne vom familiariza cu unele dintre provocările măsurătorilor sau metricilor de echitate și vom vedea ce pot însemna aceste metrici în practică.
Conținutul este organizat pe trei niveluri	**1. <u>Utilizarea Data Science pentru binele social</u>** Privind diferite cazuri de utilizare, în special „Amnesty Italy Use Case”, veți obține o imagine de ansamblu asupra modului în care Data Science poate fi utilizată în scopuri bune. **1.1 Prezentare generală a cazurilor în care se poate utiliza Data science pentru binele general** Cel mai bun mod de a înțelege impactul pozitiv pe care Data Science îl poate avea asupra oamenilor și a planetei este să analizăm câteva exemple din trecutul recent. Ritmul rapid al schimbărilor tehnologice declanșează schimbări inclusiv pe piața forței de muncă – vechile locuri de muncă și profesii dispar și sunt înlocuite de altele noi. Acest lucru are ca efect apariția șomajului în unele

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

sectoare, în timp ce în altele, angajatorilor le este greu să găsească angajați calificați. Dar, de fapt, multe competențe obținute în sectoarele „pe cale de dispariție” ar putea fi ușor adaptate și reintegrate în sectoare noi. În proiectul-pilot SkillsFuture Singapore, Data Science este utilizată pentru a detecta aceste abilități „reutilizabile” și pentru a sprijini șomerii cu cursuri de formare direcționate pentru a-și ajusta abilitățile la nevoile din sectoarele industriale în expansiune.

AI (Inteligența artificială) poate fi, de asemenea, utilizată pentru a îmbunătăți capacitatea de predicție a replicilor digitale, de exemplu pentru a contribui la creșterea rezilienței lanțului de aprovizionare. Replicile digitale folosesc datele disponibile la nivelul unei companii – fie date generate intern prin procese operaționale, tranzacționale sau alte procese, fie date disponibile public, cum ar fi monitorizarea vremii – pentru a simula lanțul de aprovizionare. Sistemele de inteligență artificială instruite prin algoritmi de învățare pot fi adăugate acestor replici digitale, permițând companiilor să exploreze efectele mai multor scenarii „ce-ar fi dacă”, cum ar fi impactul unei tornade, și să dezvolte măsuri pentru a reacționa la astfel de scenarii [2].

Sistemele AI pot fi utilizate într-o varietate de moduri în vederea atingerii obiectivelor climatice. De exemplu, Fero Labs folosește inteligența artificială pentru a ajuta producătorii de oțel să reducă utilizarea ingredientelor extrase cu până la 34%, prevenind producerea a aproximativ 450.000 de tone de emisii de CO2 pe an, în timp ce Proiectul Mapping the Andean Amazon folosește inteligența artificială pentru a monitoriza defrișarea prin intermediul imaginilor prin satelit pentru a ajuta la descoperirea defrișărilor ilegale și pentru a sprijini reacția politică [3].

Una dintre provocările asociate vehiculelor electrice este că acestea necesită acces la infrastructura electrică special concepută pentru ele – și anume stațiile de încărcare pentru mașini electrice. Dacă mai multe mașini au nevoie de aceeași infrastructură în același timp, acest lucru poate reprezenta o provocare semnificativă pentru rețeaua electrică. Astfel, unul dintre obstacolele în calea adoptării pe scară largă a surselor de energie regenerabilă este fluctuația mare a disponibilității energiei și capacitatea limitată de a stoca electricitate la orele de vârf de disponibilitate, pentru a o distribui apoi la orele de vârf de utilizare. Tehnologiile de la vehicul la rețea, care permit mașinilor electrice să fie folosite ca „depozite” pentru surplusul de energie astfel încât rețeaua să poată extragă energie din mașini atunci când mașinile nu sunt utilizate, pot ajuta la atenuarea problemei. Folosind AI, Caltech a dezvoltat un sistem de încărcare adaptivă care programează când și ce vehicul să încarce și când și câtă energie poate fi reintrodusă în rețea, pe baza orelor de plecare transmise de șofer. Acest lucru reduce stresul general pus asupra rețelei electrice și deschide o posibilitate interesantă pentru mașinile electrice de a ușura efectiv o parte din sarcina rețelelor electrice [4].

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Lanțurile de aprovizionare sunt incredibil de complexe, ceea ce reprezintă o provocare pentru legislație, cum ar fi Legea Uyghur din SUA, de Prevenire a Muncii Forțate, care urmărește să impună standarde sociale sau de mediu mai ridicate pentru produse. Atlasul Altana combină informațiile geolocalizate despre locațiile și facilitățile companiei cu datele de proprietate corporativă pentru a mapa relațiile comerciale între sectoare. Acest lucru ajută companiile să se conformeze mai eficient cu o astfel de legislație și să ia măsuri pe cont propriu împotriva unor probleme precum munca forțată [5].

Turbinele eoliene sunt o sursă importantă de energie regenerabilă, dar puterea lor depinde de un factor greu de controlat: vântul. Acest lucru reprezintă o provocare pentru rețeaua energetică, dar și pentru departamentul de vânzări al furnizorilor de energie eoliană, având în vedere că o energie mai previzibilă poate obține și prețuri mai mari. Pentru a susține cazul de afaceri al parcurilor eoliene, DeepMind a dezvoltat o rețea neuronală instruită cu ajutorul prognozele meteo și al dateler operaționale istorice, rețea care pooate prezice producția parcului eolian cu 36 de ore înainte, obținându-se astfel o valoare cu 20% mai mare pentru energia produsă [6].

1.2 Cazul Amnesty Italy

Rețelele de socializare reprezintă o parte importantă a sferei publice. Pentru a investiga modul în care se dezvoltă discursul politic privind problemele legate de drepturile omului și modul în care acesta afectează grupurile dezavantajate, Anmesty International Italia efectuează în fiecare an o monitorizare numită Barometrul Hate (Barometre dell'Odio) folosind tehnici de Data Science.

Datele sunt colectate prin intermediul API-urilor publice Facebook și Twitter, dintr-o listă de conturi și profiluri publice furnizată de Amnesty. De obicei, perioada de monitorizare este cuprinsă între patru și opt săptămâni (2021 a avut o perioadă de monitorizare extinsă de 16 săptămâni). Pentru această perioadă, comentariile din cele mai active conturi sunt eșantionate aleatoriu, totalizând un set de 30.000-100.000 de comentarii, care sunt etichetate de către voluntari instruiți de la Amnesty, în ceea ce privește subiectul și cât de ofensatoare sunt. Toate etichetele sunt verificate încrucișat, ceea ce înseamnă că fiecare comentariu este etichetat de doi până la trei voluntari și orice neconcordanțe sunt rezolvate de Consiliul pentru Ură al Amnesty (Tavolo dell'Odio).

Exemplu: Alegerile pentru Parlamentul European 2019
În perioada premergătoare alegerilor pentru Parlamentul European din 2019, profilurile publice a 461 de candidați pe Twitter și Facebook în cele șase

Co-funded by the Erasmus+ Programme of the European Union

săptămâni premergătoare alegerilor (15 aprilie – 24 mai 2019). În total, au fost colectate inițial 27.000 de postări și 4 milioane de comentarii. Într-o a doua etapă, dimensiunea setului de date a trebuit să fie redusă pentru ca voluntarii să poată gestiona setul de date, în funcție de amploarea activității pe rețelele sociale a profilurilor, asigurând în același timp reprezentarea generală a tuturor părților, a tuturor regiunilor și a cel puțin o femeie și un bărbat per partid. În acest fel, setul de date final a inclus activități de social media aferente a 77 de politicieni: 80% dintre postări au fost etichetate de 150 de voluntari Amnesty, alături de o eșantionare aleatorie de 100 de mii de comentarii.

Rezultatele [8] arată că discursul instigator la ură nu este distribuit aleatoriu, ci este concentrat în clustere. Chiar dacă prevalența sa globală pe platformele de social media este estimată la aproximativ 1%, este mai probabil să apară în legătură cu anumite grupuri și subiecte și atinge vârfuri în anumite momente. De exemplu, discursul instigator la ură este mai probabil să apară atunci când discuția implică migrație, romi, minorități religioase sau femei.

Uitându-ne cu atenție la date, se pot observa și anumite tipare. Discursul instigator la ură generează mai mult discurs instigator la ură, dar este, de asemenea, mai probabil să determine interacțiuni (cum ar fi reacții, distribuiri sau comentarii). Poate fi folosit și pentru a exclude în mod activ oamenii de pe platformele de social media: de exemplu, în timpul campaniei de monitorizare din 2020, s-a observat cum două femei au fost vizate în mod specific de discursul instigator la ură și trei au fost înlăturate de pe platformele de social media [9].

2. Data science nu face întotdeauna bine

Din păcate, la fel ca orice altă tehnologie, inteligența artificială și Data Science pot fi, de asemenea, folosite în scopuri rele sau pot avea consecințe nedorite. Cu toate acestea, spre deosebire de alte instrumente, AI automatizează deciziile pentru noi și, prin urmare, are un potențial și mai mare de a provoca rău. Prin urmare, trebuie să fim conștienți de faptul că AI și Data Science pot avea un impact negativ asupra oamenilor, societății și mediului.

2.1 Exemple majore cunoscute

Data science își propune să ne ajute să luăm decizii mai bune pe baza datelor, făcând posibilă procesarea unor cantități mari de date sau a unor tipuri foarte diverse de informații. După cum am văzut anterior, data science poate fi folosită pentru a monitoriza sau îmbunătăți procesele care ajută la transformarea lumii într-un loc mai bun. Cu toate acestea, istoria recentă ne-a arătat că nu putem avea încredere orbește în rezultatele algoritmilor, mai ales atunci când aceste rezultate pot avea un impact negativ grav asupra vieții noastre.

Exemple binecunoscute de impact negativ al AI au apărut în aplicarea inteligenței artificiale în domenii variate, de la sănătate la muncă și mediu:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

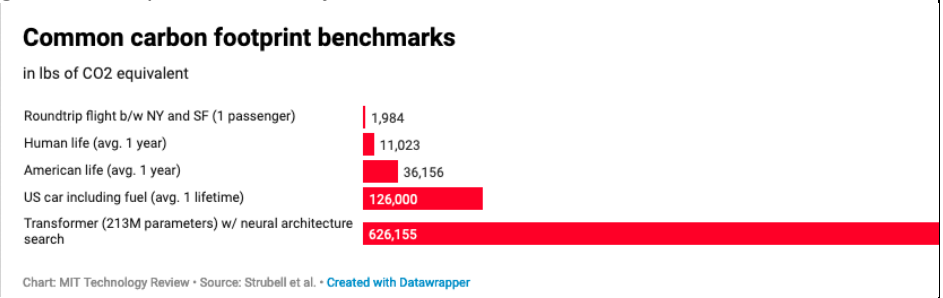
1. Spitalele din SUA se bazează acum pe algoritmi care să îi ajute să evalueze cât de bolnavi sunt pacienții, pentru a determina dacă au nevoie de îngrijire în spital sau ambulatorie. Un studiu a constatat că evaluările unui sistem foarte utilizat pe scară largă au fost denaturate într-o manieră părtinitoare din punct de vedere rasial: pacienții de culoare erau de fapt mai bolnavi decât cei albi care au primit aceeași evaluare de risc. Acest lucru s-a întâmplat probabil din cauza faptului că algoritmul a folosit costurile istorice ale sănătății ca proxy pentru nevoile de sănătate – cu toate acestea, deoarece sistemul de sănătate din SUA a fost afectat istoric de tratament inechitabil, s-au cheltuit mai puțini bani pentru a acoperi nevoile de sănătate ale pacienților de culoare. Astfel, algoritmul a concluzionat în mod greșit că aceștia sunt mai sănătoși decât pacienții albi care sunt de fapt la fel de bolnavi [10].
2. Amazon a creat un instrument de recrutare AI pentru a ajuta Departamentul de Resurse Umane să găsească personalul potrivit pentru posturile tehnice și l-a instruit cu privire la CV-urile trimise companiei în ultimii zece ani. Însă, deoarece majoritatea acestor aplicații au venit de la bărbați, Amazon și-a dat seama curând că sistemul său de recrutare nu evaluează candidații într-un mod neutru din punct de vedere al genului. Sistemul AI a penalizat CV-urile trimise de femei și care conțineau cuvinte precum „femei”. Software-ul a trebuit să fie demontat și până acum nu a fost repus în funcțiune [11].
3. În 2015, clasificatorul de imagini de la Google a etichetat o persoană de culoare drept „gorilă”. Google și-a cerut scuze, dar a optat pentru o remediere rapidă prin simpla cenzură a cuvintelor „gorilă”, „cimpanzeu” și „maimuță” din căutări și etichete de imagini. Șase ani mai târziu, Facebook a clasificat un bărbat de culoare într-un videoclip drept primată, recomandând utilizatorilor să continue să vizioneze videoclipuri cu primate. [12]

Acestea sunt doar câteva dintre exemplele care ilustrează efectele potențial negative. Data Science și inteligența artificială au nevoie de date - și adesea, aceste date sunt etichetate sau prelucrate de lucrători prost plătiți, care muncesc în condiții foarte stresante și sunt adesea expuși la conținut violent sau perturbator [13]. Algoritmii pot fi folosiți pentru a clasifica angajații sau antreprenorii într-o manieră discriminatorie și care poate duce la pierderea oportunităților [14]. Data Science și IA sunt costisitoare din punct de vedere computațional – ceea ce înseamnă că sunt, de asemenea, consumatoare de resurse; acesta este în special cazul modelelor mari care necesită multe

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

ajustări (mulți parametri), cum ar fi modelele tip ”transformer” incluse în graficul comparativ de mai jos [15].



Exercițiu: Dacă doriți să deveniți singur detectiv de imparțialitate, accesați Google Translate (sau deepl.com) și traduceți din engleză în germană:
Engleză: My doctor is clever. She immediately found the solution
Google Germană:

Engleză: My secretary is clever. He immediately found the solution
Google Germană:

Google a încercat să abordeze această problemă în 2018, după un mare protest cu privire la traducerea în roluri de gen stereotipe din limbi neutre de gen, dar după cum puteți descoperi singur, cinci ani mai târziu, problemele rămân.

2.2. Prezentare generală a principalelor riscuri

De la utilizarea roboților pentru a crea nuduri deepfake pe Telegram, generarea de avatare sexualizate ale femeilor (dar nu ale bărbaților), nedezvoltarea de funcționalități utile unui anumit grup de persoane sau subminarea identității de gen prin clasificare binară, aplicațiile Data Science pot provoca daune.

Unul dintre principalele riscuri asociate cu AI și data science este că presupunem că tehnologia în sine – ca orice alt instrument – este lipsită de judecată și eroare umană. Însă, în această teorie uităm că noi suntem cei care creăm aceste sisteme, că alegem algoritmii, selectăm datele și decidem cum să le folosim și cum trebuie să fie implementat sistemul. Prin urmare, este fundamental să înțelegem că aplicațiile data science – chiar dacă au fost create cu cele mai bune intenții – nu sunt nici obiective, nici neutre.

Reflectați la ce poate face aplicația dvs., pentru ce este utilizată, cine este inclus/exclus și cine ar putea fi afectat în moduri diferite - consecințele pot avea multiple ramificații!

În studiul lor din 2018 [15], Joy Buolamwini și Timnit Gebru au descoperit că algoritmii de clasificare de gen care utilizează recunoașterea facială, clasificau în mod frecvent greșit femeile cu pielea mai închisă decât bărbații (și femeile) cu pielea mai deschisă. Acest lucru se datora faptului că seturile de date pe

Co-funded by the Erasmus+ Programme of the European Union

care modelele investigate au fost instruite conțineau o pondere disproporționată de imagini cu bărbați și femei cu pielea deschisă la culoare.

Două studii din 2019 au arătat că algoritmii utilizați pentru a detecta discursul ofensator pe platformele online aveau mai multe șanse să clasifice tiparele de vorbire comune în rândul americanilor de culoare ca fiind ofensatoare – iar seturile de date au afișat, în mod similar, o părtinire larg răspândită față de engleza afro-americană [16]. Aceasta arată cât de importantă este etichetarea setului de date: dacă datele sunt etichetate într-un mod părtinitor, rezultatele vor fi și ele părtinitoare.

Trebuie să recunoaștem că aplicațiile Data Science nu sunt perfecte, iar erorile lor nu sunt distribuite aleatoriu: de fapt, aceste sisteme tind să eșueze mai des pentru grupurile demografice care istoric au fost marginalizate sau vulnerabile.

În plus, aplicațiile Data Science pot consuma foarte mult date, implicând diferite probleme:

- Confidențialitate: modelele AI care se bazează pe tot mai multe date stimulează colectarea de date în diferite domenii. Aceasta înseamnă că o mulțime de date ajung să fie colectate despre oameni, cu implicații importante pentru confidențialitate. De exemplu, deși uneori poate fi practic din perspectiva consumatorului să știi unde se află exact coletul tău în acest moment, iar din perspectiva unui furnizor de servicii de curierat poate fi practic să ai astfel de date pentru a optimiza rutele, urmărirea vehiculului în care este livrat coletul înseamnă și urmărirea persoanei care conduce vehiculul.
- Protecția datelor: multe dintre datele colectate vă pot permite să identificați persoane și, prin urmare, sunt considerate date de identificare personală - cum ar fi exemplul de urmărire a coletului pe care tocmai l-am discutat. Astfel de date nu numai că pot fi utilizate greșit în continuare, dar pot fi utilizate și pentru a le restricționa oportunitățile, motiv pentru care Regulamentul general al UE privind protecția datelor are o politică strictă de minimizare a datelor.
- Calitate slabă a datelor: este posibil să fi auzit de expresia „garbage in, garbage out” pentru a descrie modul în care calitatea slabă a datelor poate duce la rezultate proaste. Acest lucru înseamnă că doar având un volum mare de date nu va îmbunătăți modelul sau rezultatele. Dimpotrivă, un set mare de date care este prost etichetat, procesat incorect și care conține multe date irelevante vă va înrăutăți rezultatele. Rețineți: cea mai mare parte a timpului petrecut pe proiecte de Data Science și AI este dedicat creării unui set de date de înaltă calitate pe care apoi să îl puteți utiliza în mod fiabil și repetat. Fă ca acest efort să conteze!

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Pentru a contracara riscurile care decurg din data science și IA, au fost elaborate până în prezent peste 80 de Ghiduri de etică diferite: printre cele mai cunoscute sunt cele realizate de organizații internaționale precum OCDE, UNESCO, UNICEF; dar și de marile companii de tehnologie, precum Google și Microsoft.

Problema acestor standarde de etică este că ele nu sunt nici obligatorii din punct de vedere juridic, nici aplicabile: nu există consecințe pentru nerespectare. Standardele de etică ne ajută să stabilim direcția corectă și ne oferă îndrumări pentru ceea ce este greșit și corect, dar, caracterul voluntar al unor astfel de inițiative înseamnă că ele reprezintă doar un concept frumos, în loc să fie o necesitate.

3. AI de încredere

În această secțiune, ne vom uita la caracteristicile așa-numitului „AI de încredere", vom analiza de unde vine noțiunea și de ce este important. Ne vom concentra asupra subiectului părtinirii nedorite care poate duce la discriminare și asupra modalităților de măsurare a corectitudinii cu ajutorul unei matrice de confuzie.

3.1 AI de încredere

Uniunea Europeană și-a creat, de asemenea, propriile standarde de etică „Ethics Guidelines for Trustworthy Artificial Intelligence" [17]. Un document pregătit de Grupul de experți la nivel înalt pentru inteligența artificială (AI HLEG), un grup de experți independenți care a fost înființat de Comisia Europeană în iunie 2018, ca parte a strategiei UE în domeniul inteligenței artificiale.

HLEG UE a stabilit următoarele caracteristici ale unui sistem de AI de încredere, bazat pe Carta drepturilor fundamentale a UE:[1]

(1) agenție umană și supraveghere: sistemele AI ar trebui să fie înțelese de oameni în măsura în care deciziile lor pot fi contestate, iar oamenii ar trebui să poată interveni întotdeauna în sistemele AI

(2) robustețe tehnică și siguranță: sistemele AI ar trebui să fie capabile să facă față unei varietăți de situații pe care le-ar putea întâlni, inclusiv atacuri cibernetice și ar trebui să fie proiectate pe baze de securitate și siguranță.

(3) confidențialitate și guvernanța datelor: sistemele AI nu ar trebui să submineze dreptul nimănui la confidențialitate, persoanele vizate ar trebui să aibă control deplin asupra modului în care sunt utilizate datele lor, iar datele

[1] The Charter of Fundamental Rights of the European Union brings together the most important personal freedoms and rights enjoyed by citizens of the EU into one legally binding document. See, for example, https://fra.europa.eu/en/eu-charter

Co-funded by the Erasmus+ Programme of the European Union

nu ar trebui să fie folosite pentru a dăuna sau discrimina persoanele vizate. În plus, trebuie să existe un sistem adecvat de guvernare a datelor pentru a se asigura că setul de date este de înaltă calitate și nu poate fi accesat în scopuri nelegitime.

(4) transparență: deciziile luate de sistemele AI ar trebui să poată fi urmărite și explicate oamenilor, iar limitele sistemului AI ar trebui comunicate în mod clar

(5) diversitate, nediscriminare și corectitudine: seturile de date părtinitoare cauzează probleme, dar și modelele părtinitoare sau sistemele de inteligență artificială care au efecte disproporționate asupra anumitor grupuri (de obicei dezavantajate) sunt dăunătoare. Din acest motiv, diversitatea reprezentării și participării în toate etapele ciclului de dezvoltare a AI sunt esențiale pentru identificarea precoce a posibilelor daune și dezvoltarea mecanismelor adecvate de prevenire și atenuare.

(6) bunăstarea mediului și a societății: sistemele AI au un impact real asupra societății și asupra mediului, nu numai asupra indivizilor. Aceasta înseamnă că, în unele zone, utilizarea sistemelor AI ar trebui să fie bine reflectată și toate sistemele AI ar trebui proiectate într-o manieră durabilă din punct de vedere ecologic și social.

(7) responsabilitate: sistemele de inteligență artificială ar trebui să fie auditabile și ar trebui identificate și abordate în prealabil efectele negative potențiale, precum și compromisurile, oferind posibilitatea unei remedieri eficiente în cazul în care este cauzat un prejudiciu

În timp ce ghidul EU HLEG merge mai departe decât simplele orientări de etică, prin fundamentarea principiilor în Carta drepturilor fundamentale a UE (un cadru legal), vom vedea în secțiunea următoare, pe baza exemplului echității și nediscriminării (principiul 5), că mai este un drum lung de parcurs de la principiu până la implementare.

3.2. Prejudecăți, echitate, nediscriminare

Cu toții avem dreptul uman de a fi tratați într-un mod echitabil. Dar ce se înțelege prin corectitudine sau echitate? În esență, corectitudinea este un concept subiectiv și depinde de cultură și context. În încercarea de a ocoli această problemă dificilă, multe cercetări s-au concentrat pe problema prejudecății, a părtinirii (a așa-numitului ”bias”) în AI.

Cu toate acestea, în contextul Data Science și al învățării automate în general, multe definiții diferite ale părtinirii se întâlnesc și se pot contrazice (utilizare colocvială vs. Statistică vs. Deep Learning). Aceasta este o problemă deoarece oamenii din medii disciplinare diferite vorbesc despre părtinire (*bias*), dar de

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

fapt, pentru ei nu înseamnă același lucru. În contextul unui AI de încredere, vom considera părtinirea (*bias*) drept o prejudecată care favorizează un grup în detrimentul altuia.

Există multe tipuri diferite de prejudecăți sau erori (*bias*), cum ar fi prejudecata societală, prejudecata de confirmare, prejudecata în grup, prejudecata automată, prejudecata temporală, denaturare (bias) determinată de variabile omise, erori de eșantionare, erori de reprezentare, erori de măsurare, erori de evaluare și multe altele.

Toate aceste prejudecăți - în date, în sistemul AI sau care decurg din interacțiunea oamenilor cu prejudecăți cu sistemul AI - pot duce la un tratament inechitabil și discriminare, ceea ce înseamnă tratament nedrept sau prejudiciabil al diferitelor categorii de persoane, de exemplu, pe motive de etnie, vârstă, sex sau dizabilitate.

Dar cum să detectăm și să măsurăm părtinirea?

Primul pas este să verificați calitatea datelor dvs., care este una dintre cele mai comune modalități prin care părtinirea se strecoară în setul de date. Dar chiar dacă nu există probleme în datele dvs., modelul poate fi în continuare părtinitor.

De obicei, puteți detecta părtinirea doar ca efect asupra rezultatului modelului. Faceți acest lucru cu așa-numita metrică de echitate, care este subiectul secțiunii următoare. După cum puteți vedea, încercarea de a evita definirea echității prin analiza în schimb a părtinirii, nu a ajuns foarte departe.

3.3. Metrică de echitate

Deoarece nu există o definiție unică și perfectă a echității, nu există o singură metrică potrivită pentru a măsura echitatea și este imposibilă o soluție unică pentru toate. În schimb, există multe tipuri diferite de echitate și modalități de măsurare, incluzând: corectitudinea grupului, paritatea statistică condiționată, rata de eroare fals pozitivă, rata de eroare fals negativă, rata de acuratețe a utilizării condiționate, rata de acuratețe generală, corectitudinea testelor, calibrarea corectă, echitate prin neconștientizare, corectitudinea contrafactuală și multe altele.

Din păcate, nu le puteți testa pur și simplu pe toate pentru a vă asigura că algoritmul dvs. este corect, deoarece aceste valori pot duce la rezultate contradictorii. De exemplu, este imposibil din punct de vedere matematic să se îndeplinească cerințele atât pentru paritatea predictivă, cât și pentru cotele egalizate. Luați în considerare următoarea formulă, determinată în [18]:

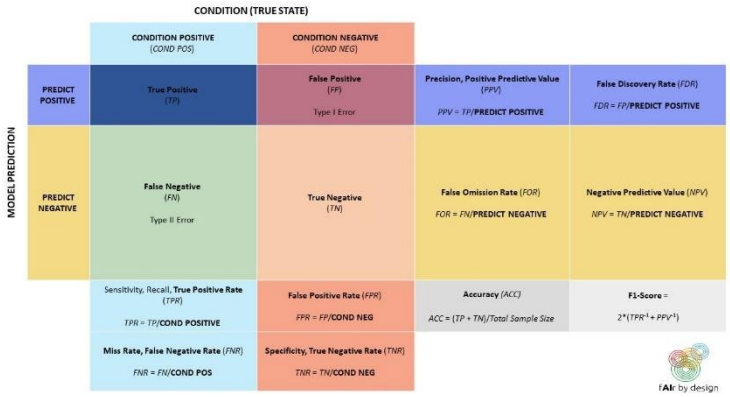
$$\mathbf{FPR = (1 - FNR)\ \frac{p}{1-p}.\frac{1 - PPV}{PPV}}$$

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

p din formulă se referă la prevalența clasei POZITIV și puteți folosi matricea de confuzie de mai jos pentru a înțelege ceilalți termeni. Acum să presupunem că aveți două grupuri demografice, G1 și G2, cu prevalența p1 și p2. Dacă cotele egalizate sunt valabile, atunci FPR și FNR sunt aceleași pentru ambele grupuri. Dacă paritatea predictivă este valabilă, atunci și PPV este același pentru ambele grupuri. Introducând toate aceste informații în formula de mai sus, veți avea cu două ecuații, una pentru G1 și una pentru G2. Aplicând câteva noțiuni de aglebră, veti obține că și p1 și p2 **trebuie** să fie egale.

Pentru a rezuma: dacă atât cotele egalizate, cât și paritatea predictivă sunt adevărate, atunci prevalența trebuie să fie aceeași pentru ambele grupuri. În schimb, dacă prevalența nu este aceeași pentru ambele grupuri, atunci cotele egalizate și paritatea predictivă **nu pot fi** ambele adevărate!



Imposibilitatea matematică de a satisface toate metricele de echitate simultan înseamnă că trebuie luată o decizie cu privire la definiția care trebuie aplicată echității. Din păcate, în prezent nu există un cadru legal sau exemple de bune practici – și aceasta înseamnă că trebuie să luați în considerare cu atenție contextul aplicației dvs. de inteligență artificială înainte de a alege metrica pentru evaluarea impactului acesteia în termeni de echitate.

Pentru a înțelege implicațiile de a avea mai multe definiții ale echității care nu sunt compatibile și importanța de a conveni asupra unei definiții înainte de implementarea unor astfel de sisteme, vom arunca o privire asupra unui exemplu din viața reală care a declanșat o mare parte a cercetării și a dezbaterii asupra prejudecății algoritmilor în comunitatea data science și ML. COMPAS este un sistem AI dezvoltat de o companie numită Northpointe și este utilizat în sistemul de justiție penală al Statelor Unite pentru a estima riscul de recidivă al inculpatului (cu alte cuvinte, pentru a evalua riscul inculpatului de a comite o altă infracțiune în viitor). Acest scor de risc este apoi utilizat pentru a lua decizii cu privire la eliberare condiționată sau eliberare anticipată.

Pentru a funcționa, sistemul AI s-a bazat pe date istorice ale criminalității, care se refereau la infractorii din trecut și dacă aceștia au fost arestați din

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

nou pentru o altă infracțiune după eliberare - deci conținea informații referitoare la tipurile de inculpați susceptibili să comită din nou infracțiuni (și să fie prinși făcând acest lucru!). Aceste înregistrări au fost folosite pentru a instrui modelul pentru a prezice riscul de recidivă al inculpaților care nu au făcut parte din setul de date, odată ce sistemul a intrat în funcțiune. Aceasta înseamnă că probabilitatea de recidivă pentru fiecare inculpat a fost calculată și apoi inculpații au fost clasificați cu risc scăzut sau cu risc ridicat.

În mai 2016, ProPublica a publicat un articol în care se indică faptul că predicțiile acestui model de modelare a recidivei erau părtinitoare [18; vezi, de asemenea, 19, 20]: ProPublica a demonstrat că formula sistemului de inteligență artificială era foarte probabil să semnaleze în mod fals inculpații de culoare ca fiind cu risc ridicat de recidivă, etichetându-i în mod greșit în acest fel la aproape dublul ratei față de inculpații albi (42% față de 22%); în același timp, inculpații albi au fost etichetați greșit ca fiind cu risc scăzut mai des decât inculpații de culoare. [2]

Dacă ne uităm la matricea de confuzie de mai sus, putem observa că ProPublica spunea că COMPAS a fost nedreaptă pentru că FPR și FNR nu erau la fel pentru inculpații negri vs. inculpații albi. Aceasta este chiar metrica cotelor egalizate:

1. Cotele egalizate

Cotele egalizate înseamnă că, în cadrul fiecărei categorii reale de risc, procentul de predicții fals negative și de predicții fals pozitive este egal pentru fiecare grup demografic. Întrebarea nu se mai concentrează pe acuratețea generală a modelului, ci mai degrabă descompune tipurile de erori pe care modelul le poate face (fals pozitive și fals negative) și necesită ca erorile modelului să fie comparabile: FPR este egal între grupuri și FNR este egal între grupuri.

Northpointe și-a apărat sistemul COMPAS împotriva acuzației de părtinire, subliniind că, dacă modelul prevedea că un inculpat prezintă un risc ridicat, atunci șansa ca acesta să recidiveze efectiv era aceeași, indiferent de grupul demografic de care aparținea inculpatul. Northpointe spune: probabilitatea unui caz pozitiv real, când modelul prezice un caz pozitiv, este aceeași pentru toate grupurile. Aceasta este cunoscută sub denumirea de metrica de echitate Paritate Predictivă.

2. Paritate Predictivă

Paritatea predictivă înseamnă că proporția de inculpați cu risc ridicat prezis corect este aceeași, indiferent de criteriile demografice. Cu alte cuvinte, paritatea predictivă se referă la conceptul din ML și AI conform căruia modelele predictive utilizate ar trebui să producă aceeași valoare predictivă pozitivă (PPV) pentru diferite grupuri, indiferent de apartenența acestora la o clasă protejată (de exemplu, rasă, sex, vârstă etc.). PPV este o metrică utilizată pentru a evalua proporția de predicții pozitive adevărate (instanțele pozitive clasificate corect) dintre toate cazurile în care modelul a prezis

[2] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Co-funded by the Erasmus+ Programme of the European Union

pozitiv. Cu toate acestea, o astfel de măsurătoare nu ține cont de prevalența totală a instanțelor dintr-un set de date!
Reformulând, Paritatea Predictivă ia în considerare echitatea analizând erorile în raport cu clasa *prezisă*, în timp ce cotele egalizate analizează erorile în raport cu clasa *reală*. Dacă este mai important să optimizați PPV (și, prin urmare, ați prefera echitatea predictivă a parității) sau dacă preferați să minimizați FPR (și, prin urmare, preferați cotele egalizate) ține foarte mult de perspectivă. De exemplu, ce măsură de eroare este mai importantă pentru dvs. dacă ați primit un diagnostic medical de la un sistem AI? Și ce măsură de eroare este mai importantă într-un algoritm de angajare folosit pentru a angaja pentru un loc de muncă pentru care ai aplicat? Vă puteți gândi la situații în care ați putea considera PPV mai important și alte situații în care ați prefera un FPR scăzut?
Dacă doriți să aflați mai multe despre diferitele definiții ale echității (de fapt, în prezent există mai mult de 21), despre cum să le măsurați și despre diferențele dintre ele, consultați „Explicarea definițiilor echității" [22].
Reflectați: Revenind la exemplul COMPAS, ce definiție ați considera că este cea echitabilă?
Explicați: Este posibil să se satisfacă ambele definiții de echitate?
Răspuns: Trebuie să înțelegem prevalența recidivei. În SUA, rata generală de recidivă pentru inculpații de culoare este mai mare decât pentru inculpații albi (52% față de 39%). Conform formulei pe care am văzut-o mai sus, aceasta înseamnă că nu este posibil ca ambele definiții ale echității să fie adevărate.
Cazul COMPAS exemplifică modul în care problemele sociale au un impact asupra datelor care sunt disponibile. Forțele de poliție alocate suplimentar comunităților de culoare înseamnă că probabilitatea de arestări efectuate sau de incidente înregistrate este mai mare pentru aceste comunități. Prin urmare, datele părtinitoare sunt introduse în modele. În plus – asta înseamnă că rata de recidivă percepută pentru cele două populații nu este aceeași, forțând decizii dificile cu privire la metrica de echitate folosită – adică ceea ce este cu adevărat echitabil în acest context.
Problema reală este că există părtiniri sistemice în sistemul judiciar (în SUA, dar și în alte părți), care nu pot fi rezolvate pur și simplu prin introducerea mai multor date (cazuri istorice) în sistem. O discuție excelentă a problemelor legate de utilizarea datelor nepotrivite pentru a determina predicții în poliție poate fi găsită în "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice" [22].
Prejudecățile sistemice afectează și alte domenii de aplicare, fie că se referă la sănătate, educație sau cât plătiți pentru produse sau servicii. Uneori, putem alege instrumentele potrivite pentru a ține cont de astfel de părtiniri sistemice. Alteori, trebuie să recunoaștem că condițiile nu sunt potrivite pentru o utilizare sigură a algoritmilor. Dar, aceste alegeri nu ar trebui lăsate doar la latitudinea analistului de date, ci ar trebui să implice o multitudine de

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	părți interesate și multe expertize diferite - inclusiv, de exemplu, experți în sociologie, psihologie, drept și în domenii specifice contextului. Inteligența artificială și știința datelor nu pot face miracole și nu pot rezolva problemele societății, dar putem folosi tehnologia ca instrument pentru a scoate la lumină aceste probleme sistemice și pentru a le aborda. Pentru că „AI funcționează doar dacă funcționează pentru noi toți”[24]. 4. Concluzie Deci, să recapitulăm ceea ce am învățat: Pe de o parte, data science și AI au o mare varietate de aplicații cu impact social pozitiv. De exemplu, data science este utilă pentru a investiga modul în care rețelele sociale influențează drepturile omului. Pe de altă parte, data science și aplicațiile AI implică, de asemenea, riscuri pentru sănătate, siguranță, mediu și drepturile omului. Prejudecățile și discriminarea, preocupările legate de confidențialitate și efectele nocive asupra mediului sunt doar câteva dintre efectele posibile. Echitatea rezultatelor în data science și aplicațiile AI poate fi măsurată în diferite moduri. Construirea de aplicații AI de încredere necesită o colaborare interdisciplinară intensă: asigurându-ne că procesele noastre de dezvoltare sunt incluzive și permit o participare largă, putem construi aplicații mai bune.
Auto-evaluare (întrebări cu variante multiple și răspunsuri)	1. Numiți trei cazuri de utilizare a data science pentru binele public **A) încărcare adaptivă** **B) potrivirea aptitudinilor** **C) monitorizarea social media pentru a observa impactul asupra drepturilor omului** 2. Care dintre cele de mai jos **nu** este unul din principiile HLEG pentru AI de încredere? A) Robustețe **B) Reproductibilitate** C) Transparență 3. Metrica de echitate a Cotelor Egalizate necesită ca: A) TPR să fie egale pentru toate grupurile demografice B) FPR să fie egale pentru toate grupurile demografice **C) Ambele variante.**
Resurse (video, link-uri)	- [1] Skills adjacency detection and targeted training of missing skills: SkillsFuture Singapore, https://www.skillsfuture.gov.sg/AboutSkillsFuture - [2] AI & digital twins - simulating and practicing for resilience in the supply chain: https://www.technologyreview.com/2021/10/26/1038643/ai-reinforcement-learning-digital-twins-can-solve-supply-chain-shortages-and-save-christmas/ - [3] Reducing the footprint of recycled steel: Fero Labs uses AI to help steel manufacturers reduce the use of mined ingredients by up to 34%, preventing an estimated 450,000 tons of CO2 emissions per year: https://gpai.ai/projects/responsible-ai/environment/climate-change-and-ai.pdf - [4] Adaptive charging breaks down barriers to electric vehicle adoption. Bi-directional charging & Vehicle to Grid technologies need smart scheduling algorithms. https://ev.caltech.edu/info - [5] Using AI to detect forced labor in the supply chain: https://www.altana.ai/blog/illuminating-xinjiang-forced-labor-ecosystem - [6] Machine learning can boost the value of wind energy: https://www.deepmind.com/blog/machine-learning-can-boost-the-value-of-wind-energy

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	- [7] Barometre dell'Odio:. https://www.amnesty.it/campagne/contrasto-allhate-speech-online/ - [8] Barometre dell'Odio: Elezioni europee. https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2020/01/Amnesty-barometro-odio-2019.pdf - [9] Barometre dell'Odio: sessimo da tastiera. https://www.amnesty.it/barometro-dellodio-sessismo-da-tastiera/#sintesi - [10] Ziad Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. https://science.sciencemag.org/content/366/6464/447 - [11] The Guardian, Amazon ditched AI recruiting tool that favored men for technical jobs, October, 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine - [12] After Google's Gorillas comes Facebook's Primates: Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men, September 2021. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html - [13] - [14] - [15] Joy Buolamwini & Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf - [16] The algorithms that detect hate speech online are biased against Black people. August 2019. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter - [17] EU HLEG Guidelines for trustworthy AI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai - [18] Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data. 2017 Jun;5(2):153-163. - [19] Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing - [20] A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. October 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ - [21] Julia Dressl and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. January 2018. https://www.science.org/doi/10.1126/sciadv.aao5580 - [22] Sahil Verma, Julia Rubin: „**Fairness Definitions Explained",** 2018 ACM/IEEE International Workshop on Software Fairness; https://dl.acm.org/doi/10.1145/3194770.3194776 - [23] Richardson, R. et al, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice"; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423 - [24] D. Raji, "How our data encodes systematic racism", MIT Technology Review. https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/
Materiale adiționale	
PPT	
Bibliografie	
Realizat de	[Women in AI Austria]

Co-funded by the Erasmus+ Programme of the European Union