Algoritmi disuguali in un mondo disuguale

Quelle distorsioni che originano da dati distorti ma che sono evitabili

Erano gli anni settanta quando negli Stati Uniti furono chiuse le ultime *poorhouse*: gli istituti pubblici deputati alla reclusione dei poveri, passate alla storia per i tremendi episodi di sfruttamento e di violenza che si verificavano al loro interno. Oggi, fortunatamente, di queste strutture non resta che un lontano ricordo, ma le discriminazioni a danno delle classi più deboli continuano a essere un problema. Di recente sono state descritte diverse situazioni in cui un aumento delle disuguaglianze sociali è stato causato dall'implementazione in ambito sanitario di alcune delle più innovative e sofisticate tecnologie attualmente disponibili: gli algoritmi di machine learning. Pur essendo sviluppati per ottimizzare l'assistenza, infatti, può accadere che l'utilizzo di questi strumenti finisca per intrappolare le fasce più vulnerabili della popolazione in vere e proprie *poorhouse* digitali[1].

Dimmi il tuo codice postale e ti dirò chi sei

Per una struttura ospedaliera, si sa, i ricoveri rappresentano una delle maggiori voci di spesa. Inoltre, per i pazienti sono generalmente un'esperienza molto sgradevole, oltre che pericolosa. Per queste ragioni nel 2017 un gruppo di data scientists del Center for healthcare delivery science and innovation dell'University of Chicago medicine ha sviluppato un algoritmo di machine learning finalizzato a predire la durata dei ricoveri effettuati presso il sistema ospedaliero accademico[2]. "Volevamo assicurarci che i medici e gli operatori sapessero quali pazienti avrebbero lasciato l'ospedale entro 40 ore – spiega John Fahrenbach, uno dei data scientist che ha lavorato allo sviluppo dell'algoritmo – in modo da permettere loro di gestirli in modo prioritario rispetto a quelli che sarebbero rimasti per una settimana".

Dopo aver preso in considerazione i dati clinici dei pazienti che venivano trattati presso la loro struttura, i ricercatori hanno però cominciato a mettere in relazione i ricoveri con le informazioni demografiche. Con loro sorpresa, dai dati emergeva che il fattore in grado di predire con maggiore precisione la durata delle ospedalizzazioni era il codice postale. "Avevamo questo errore nell'algoritmo – sostiene Fahrenbach – e dovevamo capire quale fosse la causa". Approfondendo la questione i ricercatori si sono trovati di fronte a una situazione molto delicata da un punto di vista etico: i codici postali associati ai ricoveri più lunghi erano quelli relativi a quartieri abitati principalmente da poveri e afroamericani.

"A parità di condizioni mediche, le persone che provengono da aree con meno risorse impiegano più tempo per essere dimesse, perché hanno bisogno di maggiore supporto sociale – sottolinea il data scientist –, c'è chi ha bisogno di aspettare che gli venga fornita una sedia a rotelle, ci sono madri con figli al seguito che chiedono di restare una notte in più perché la situazione nelle loro case è critica". Se gli sviluppatori avessero deciso di modellare l'algoritmo esclusivamente in termini di efficienza, quindi, avrebbero finito per penalizzare i cittadini con maggiori necessità di supporto sociale, dirigendo invece le risorse verso quelli provenienti dai quartieri più benestanti.

> *Il primo passo è diventare consapevoli del problema, solo così è possibile cercare di gestirlo, di affrontarlo.* — **John Fahrenbach**

"Non sapevo cosa fare e quindi mi sono rivolto al Diversity and equity commitee della nostra università – spiega Fahrenbach –, dove ho trovato supporto. Non è semplice trattare i bias che riguardano questioni cliniche e sociali perché si tratta di variabili in relazione tra loro: sappiamo per esempio che la provenienza è associata al reddito, che è a sua volta associato al tipo di assicurazione sanitaria, ma sappiamo anche che malattie come l'anemia falciforme sono più frequenti in alcuni gruppi etnici".

Grazie al lavoro congiunto del Center for healthcare delivery science and innovation e del Diversity and equity commitee è stato però possibile intervenire per tempo e sistemare le falle del sistema prima che questo diventasse operativo. Inoltre, la collaborazione tra i due dipartimenti ha anche portato allo sviluppo di una checklist di raccomandazioni utili a garantire l'equità degli algoritmi di machine learning[3]. "Il primo passo è diventare consapevoli del problema – afferma Fahrenbach –, solo così è possibile cercare di gestirlo, di affrontarlo".

"Gli sviluppatori e gli utilizzatori degli algoritmi di intelligenza artificiale dovrebbero verificare in tutte le fasi di creazione e implementazione di queste tecnologie che i criteri di equità vengano rispettati", sottolinea Marshall Chin, docente di etica sanitaria dell'University of Chicago medicine, tra gli autori della checklist. "Bisognerebbe sempre porsi alcune domande: l'obiettivo dell'algoritmo è migliorare gli outcome sanitari per tutti i pazienti o è quello, potenzialmente problematico, di tagliare i costi? L'algoritmo è stato sviluppato con dati distorti? I clinici e gli amministratori utilizzano gli algoritmi per migliorare l'assistenza a tutti i pazienti o ne discriminano alcuni?".

Machine learning: non è razzista, ma…

Se non ci si rende conto per tempo della presenza di bias le conseguenze possono essere drammatiche. Lo dimostra uno studio, pubblicato di recente su *Science*, che ha indagato la presenza di discriminazioni razziali in un algoritmo di machine learning finalizzato a individuare i casi con maggiore necessità di assistenza sanitaria complessa[4]. Una tecnologia, questa, che negli Stati Uniti è già applicata su un bacino di circa 200 milioni di cittadini.

Prendendo in considerazione i dati relativi a 6.079 pazienti che si erano autodefiniti "neri" e 43.539 pazienti che si erano autodefiniti "bianchi", i ricercatori hanno messo in relazione lo stato di salute dei partecipanti – definitivo dal numero di *a p.28* →

I pazienti che avevano ricevuto maggiore assistenza erano gestiti in modo prioritario rispetto a quelli che ne avevano ricevuta di meno. — **Ziad Obermeyer**

da p.27 → condizioni croniche attive – con la valutazione del rischio effettuata dall'algoritmo. Dai risultati è emerso che a parità di rischio predetto i pazienti di colore avevano uno stato di salute peggiore: al 97esimo percentile di rischio, valore utilizzato come livello soglia per l'assegnazione automatica a un programma di assistenza speciale, i pazienti "neri" avevano il 26,3 per cento di condizioni croniche in più rispetto ai "bianchi".

Il risultato, come sottolineano gli autori, è una disparità sostanziale e quantificabile nel reclutamento nei programmi di assistenza speciale. Immaginando di utilizzare un algoritmo privo di questo bias razziale, infatti, i ricercatori hanno dimostrato che al 97esimo percentile di rischio la percentuale di pazienti di colore a cui venivano automaticamente assegnate risorse aggiuntive saliva dal 17,7 per cento al 46,5 per cento. "In altre parole l'algoritmo originale rinforzava le disuguaglianze – racconta Ziad Obermeyer, docente di Health policy and management della Berkley university, in California, primo autore dello studio –, i pazienti che avevano ricevuto maggiore assistenza erano gestiti in modo prioritario rispetto a quelli che ne avevano ricevuta di meno".

Disponendo anche dei dati grezzi utilizzati per lo sviluppo dell'algoritmo, gli autori hanno poi potuto indagare i meccanismi sottostanti questo bias. È emerso che per valutare i bisogni di salute di un paziente gli sviluppatori avevano preso in considerazione il totale delle spese sanitarie associate a quel paziente in un anno. Di conseguenza, l'algoritmo non prediceva la probabilità di un paziente di avere bisogno di cure aggiuntive ma bensì l'ammontare dei costi che avrebbe prodotto. Facendo ulteriori approfondimenti Obermeyer e colleghi hanno quindi scoperto che la disparità di valutazione del rischio dipendeva da un dato economico reale: i pazienti "neri" generavano costi minori di quelli "bianchi". Nello specifico, a parità di condizioni sanitarie per un paziente di colore si spendevano in media 1.801 dollari in meno all'anno.

Per cercare di risolvere il bias presente nell'algoritmo i ricercatori hanno quindi iniziato una collaborazione con l'azienda produttrice. "Tutti, sia nel settore pubblico che in quello privato, stanno cominciando a capire solo ora come questi bias si insinuano negli algoritmi", spiega Obermeyer. "La società che aveva sviluppato il software si è dimostrata molto ricettiva nei confronti della nostra ricerca e disponibile a integrarne i risultati nello sviluppo dei loro algoritmi". Utilizzando una misura combinata di predittori clinici e economici al posto dei soli costi stimati, ricercatori e sviluppatori sono infine riusciti a ridurre dell'84 per cento le assegnazioni errate legate al bias. "L'algoritmo rivisto – sottolinea con orgoglio il ricercatore – si basa su una predizione di salute e mette in primo piano le persone più bisognose, a prescindere da quanto costano".

Checklist

Come garantire l'equità negli algoritmi?

DISEGNO

- ☐ **Determinare l'obiettivo del modello e discuterlo con i diversi portatori di interesse.**
- ☐ **Assicurarsi che il modello rispetti le esigenze dei pazienti e che possa essere integrato nel workflow clinico.**
- ☐ **Discutere eventuali preoccupazioni etiche.**
- ☐ **Stabilire quali gruppi classificare come protetti.**
- ☐ **Verificare la presenza di disuguaglianze nel database di partenza.**

RACCOLTA DATI

- ☐ **Raccogliere e documentare i dati utili a costruire il modello.**
- ☐ **Assicurarsi che i dati relativi ai pazienti appartenenti a gruppi protetti siano riconoscibili (assicurandosi di non violare la privacy).**
- ☐ **Valutare che i pazienti appartenenti a gruppi protetti siano adeguatamente rappresentati.**

TRAINING DELL'ALGORITMO

- ☐ **Addestrare l'algoritmo tenendo conto degli obiettivi di equità.**

VALUTAZIONE

- ☐ **Monitorare le metriche intergruppo.**
- ☐ **Confrontare i dati relativi alla fase di distribuzione con quelli della fase di training.**
- ☐ **Valutare l'utilità dell'algoritmo in una fase iniziale senza pazienti.**

LANCIO

- ☐ **Valutare la possibilità di lanciare l'algoritmo in presenza di tutti i portatori di interesse.**

DISTRIBUZIONE

- ☐ **Monitorare i dati e le metriche durante tutte le fasi di distribuzione.**
- ☐ **Lanciare l'algoritmo in modo graduale e implementare degli alert automatici per monitorare le metriche.**
- ☐ **Prendere in considerazione la possibilità di realizzare un trial clinico formale per valutare gli outcomes per i pazienti.**
- ☐ **Raccogliere periodicamente feedback dai medici e dai pazienti.**

Fonte: Rajkomar et al. 2018[3].

Non sono gli algoritmi, siamo noi

Non sempre, tuttavia, i bias associati a un aumento delle disuguaglianze nascono in fase di sviluppo dell'algoritmo. Può accadere infatti che un sistema automatico di per sé efficiente e funzionale penalizzi un dato gruppo sociale o ne favorisca un altro per via indiretta. È quanto si è verificato, per esempio, con un algoritmo utilizzato a Los Angeles per aiutare i senza tetto a ottenere un'abitazione. Tenendo conto di diverse variabili di carattere clinico e demografico raccolte durante un colloquio, questo software assegna un punteggio di rischio (compreso tra 1 e 17) a ogni individuo che entra nel programma di assistenza domiciliare. I soggetti con un punteggio che si avvicina a 17 sono quelli per cui la vita in strada rappresenta un rischio potenzialmente letale e vengono quindi assegnati a una casa in modo prioritario. Tuttavia, in aree particolarmente povere e prive di alloggi come South Los Angeles il punteggio di rischio assegnato può rivelarsi un'arma a doppio taglio: se da un lato un individuo con punteggio pari a 16 o 17 rappresenta un candidato ideale per l'assegnazione urgente di una casa, dall'altra è possibile che le istituzioni cittadine lo ritengano – proprio sulla base di quel punteggio – troppo compromesso per gestire un'abitazione in modo autonomo.

Senza un'attenzione adeguata alla tematica dell'equità, quindi, c'è il rischio che un intervento finalizzato a migliorare la qualità della vita dei più deboli si trasformi in un censimento non autorizzato o, peggio, in una sorta di prigione digitale. Secondo Marshall Chin per evitare che ciò accada è fondamentale favorire un processo di partecipazione condiviso che tenga conto dei punti di vista di tutti gli stakeholder interessati dall'implementazione di quella tecnologia, a partire dai gruppi sociali più marginalizzati. "Nell'ambito di progetti di sviluppo di algoritmi, chi finanzia il mondo della ricerca dovrebbe sempre tenere conto di eventuali problematiche legate all'equità", conclude il docente dell'University of Chicago medicine.

Dello stesso parere è anche Obermeyer, secondo cui i finanziamenti della ricerca dovrebbero essere indirizzati allo stesso tempo verso interventi finalizzati a ridurre le disuguaglianze e verso lo sviluppo di algoritmi più efficienti: "È importante ricordare che le distorsioni originano da dati distorti, ma non sono inevitabili. Nel nostro caso non abbiamo risolto il bias nei dati, e figuriamoci nella società, ma abbiamo dimostrato che è possibile ridurre gli effetti di queste distorsioni facendo migliori scelte tecniche nella costruzione degli algoritmi".

Fabio Ambrosino

1. Eubanks V. Aumating inequality. How high-tech tools profile, prolice, and punish the poor. New York: St. Martin's Press, 2017.
2. Nordling L. Mind the gap. Nature 2019;573:S103
3. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866-72.
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.