UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea

Bias negli algoritmi di Machine Learning

Il caso degli algoritmi per l'assistenza sanitaria negli Stati Uniti

Relatori
prof. Gian Antonio Susto
dott. Alessandro Fabris

Candidato
Francesco Pio Monaco
matricola 1223746

Anno accademico 2021-2022 Settembre 2022

Sommario

L'intelligenza artificiale (AI) e gli algoritmi di machine learning (ML) sono ampiamente utilizzati nel settore pubblico e nel privato per aiutare in decisioni chiave. Molti studi hanno riportato preoccupazioni sul piano etico e sulla poca trasparenza dei dati utilizzati da algoritmi ML; questo elaborato analizza i metodi di misurazione, le cause del *bias* e le contromisure per evitarlo del lavoro di Obermeyer et al. (2019) su un algoritmo utilizzato a livello sanitario per il calcolo del livello di rischio di pazienti.

Indice

Introduzione

Negli ultimi anni assistiamo ad una crescita costante del mercato delle intelligenze artificiali (AI), sostenuta parimenti da una crescita quasi-esponenziale delle pubblicazioni scientifiche riguardo l'argomento [1].

Dalla fotografia [4] alla predizione del testo [6], le AI si diffondono nel pubblico e nel privato. Una sotto-categoria notevole delle AI è quella del Machine Learning (ML), questi algoritmi sono utilizzati con successo in contesti sanitari per rilevare melanomi maligni [5], oppure per aiutare i medici nelle prescrizioni [9].

Nati come strumenti neutrali si sono ben presto rivelati iniqui, numerosi sono gli studi che rilevano come alcuni algoritmi ML perpetuino e rinforzino *bias* dannosi, alcuni esempi sono: COMPAS [10], utilizzato per la valutazione del rischio di recidiva, il quale prevedeva con una percentuale maggiore che gli imputati neri fossero a più alto rischio di recidiva rispetto a imputati bianchi; o algoritmi usati per la valutazione della qualità dell'insegnamento (UK GCSE and A-Level grading) [11] che assegnavano sistematicamente punteggi più bassi a docenti delle scuole pubbliche.

Si delinea quindi l'importanza dell'esistenza di condizioni di *fairness* affinché non vengano "codificati" *bias* negli algoritmi.

Questo elaborato ha come obiettivo l'analisi dello studio effettuato su un algoritmo utilizzato per gestire la salute pubblica [2].

Nel primo capitolo vengono fornite delle definizioni di *fairness* e *bias*. La trattazione degli algoritmi nella sanità pubblica è affidata al secondo capitolo che si concentra, inoltre, anche sui metodi di misurazione del *bias* e i risultati ottenuti nel

caso [2]. Nel terzo capitolo ci si occupa dell'origine del *bias*. Mentre nel quarto vengono trattati i metodi per la mitigazione del *bias*. Le conclusioni, infine, traggono le fila richiamando i risultati ottenuti in letteratura.

Capitolo 1

Definizioni

1.1 Cosa intendiamo per *fairness* e *bias*

Ad oggi la letteratura offre più di 21 definizioni diverse di *fairness* [7], viene fornita ora una definizione generale, successivamente sono riportate le interpretazioni probabilistiche più utilizzate.

- *Fairness:* trattamento imparziale e giusto che non privilegi o discrimini gruppi di persone.[1]

- *Bias:* Etimologicamente proveniente dal francese *biais*, una boccia che possedeva un lato più pesante dell'altro causando una curvatura durante il rotolamento.[2] In ambito informatico indica un errore sistematico in un sistema informatico che crea risultati scorretti, come privilegiare un gruppo di persone rispetto ad un altro [8].

[1] Oxford Languages

[2] Online Etymology Dictionary

1.2 Alcune definizioni di *fairness*

Tra le definizioni più utilizzate vengono riportate ora quelle più adatte al caso che viene analizzato in seguito.

1.2.1 Equità individuale

È una metrica che lavora su coppie di individui e prevede che individui "simili" siano classificati in maniera simile. Il concetto di simile è astratto pertanto viene descritto attraverso una funzione: per una decisione binaria la funzione similarità è definita come la distanza tra due individui i, j in funzione dello spazio risultati stimati Y_i, Y_j e dei vettori delle caratteristiche $\overrightarrow{X_i}$, $\overrightarrow{X_j}$. Gli individui possiedono attributo protetto[3] X_i, X_j e un vettore di caratteristiche $\overrightarrow{v_i}$, $\overrightarrow{v_j}$; tali che $\overrightarrow{X_i} = X_i + \overrightarrow{v_i}$, $\overrightarrow{X_j} = X_j + \overrightarrow{v_j}$. Allora l'equità individuale è verificata se e solo se $d_r \leq d_{i,j}$ con d_r distanza tra i risultati stimati e $d_{i,j}$ distanza tra i vettori delle caratteristiche [12].

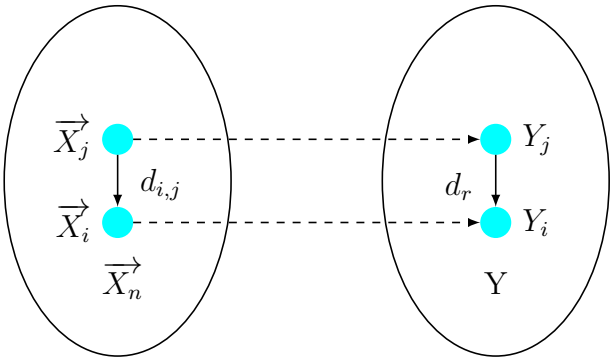

Figura 1.1. **Rappresentazione insiemistica del teorema.**

[3]Per attributo protetto intendiamo sesso, etnia o qualunque caratteristica che indichi esplicitamente l'appartenenza a un gruppo protetto.

1.2.2 Equità di gruppo

Definita in letteratura anche come equità statistica, lavora su gruppi al posto di individui singoli. I gruppi sono definiti tramite attributi protetti e la metrica richiede che l'algoritmo abbia performance simili attraverso i diversi gruppi; viene rappresentata attraverso diversi postulati, talvolta incompatibili tra loro.

Postulati come Parità demografica (Postulato 1) e Bilanciamento del tasso di errori (Postulato 2) lavorano sulla distribuzione del risultato predetto in funzione del risultato reale, mentre postulati come Calibrazione (Postulato 3) lavorano sulla distribuzione del risultato reale in funzione del risultato predetto.

Sono riportati ora i dati utilizzati dai vari postulati per una decisione binaria e tre postulati per verificare l'equità di gruppo [14]:

Siano date due classi $Y = \{a, b\}$ a classe positiva, b negativa, alle quali assegnare n campioni, che possiedono caratteristica protetta $A = \{0, 1\}$, attraverso un punteggio R che è assegnato dall'algoritmo, definiamo

- *Falsi Positivi(FP)/Negativi(FN):* Le predizioni errate delle classi, campioni negativi reali classificati come positivi e viceversa.
- *Veri Positivi(TP)/Negativi(TN):* Le predizioni corrette delle classi, campioni positivi sia reali che predetti e viceversa.
- *Tasso di Falsi Positivi(FTR)/Negativi(FNR):* Proporzione di campioni positivi/negativi per i quali è stata predetta la classe errata.
- *Recall:* Percentuale di positivi selezionati su tutti i positivi reali.

1. *Parità demografica:* Secondo questo postulato la percentuale di positivi (appartenenti alla classe a) tra i gruppi della caratteristica protetta deve essere simile. Gruppi diversi devono avere in media classificazioni simili [13] e soddisfare quindi questa equazione:

$$\mathrm{P}[Y = a|A = 1] = \mathrm{P}[Y = a|A = 0]$$

2. *Bilanciamento del tasso di errori:* Per questo postulato FNR e FTR devono essere simili tra diversi gruppi protetti. Gruppi diversi devono avere in media una percentuale di errori simile [14] devono quindi soddisfare le seguenti equazioni:

$$\mathrm{P}[\hat{Y} = a|Y = b, A = 0] = \mathrm{P}[\hat{Y} = a|Y = b, A = 1] \tag{1.1}$$

$$\mathrm{P}[\hat{Y} = b|Y = a, A = 0] = \mathrm{P}[\hat{Y} = b|Y = a, A = 1] \tag{1.2}$$

3. *Calibrazione:* Al contrario degli altri postulati la calibrazione lavora sul punteggio R; dato un $R = r_1$ è richiesto che la proporzione di campioni nella classe positiva sia la stessa tra i gruppi protetti, soddisfacendo la seguente equazione:

$$\mathrm{P}[Y = a|R = r_1, A = 0] = \mathrm{P}[Y = a|R = r_1, A = 1], \forall r_1 \in R$$

1.2.3 Equità AUC

Illustra la probabilità che dati due campioni i, j casuali l'algoritmo riesca attraverso il punteggio R a classificarli correttamente. Vengono calcolati per i diversi gruppi le aree sotto la curva ROC (AUC), dove ROC è una funzione delle coppie (FTR, TNR); $AUC \in [0,1]$, dove 0 indica che l'algoritmo prevede esattamente il contrario ed 1 predizioni perfette; comparando le diverse AUC si ottengono informazioni sull'equità dell'algoritmo, come mostrato in (Figura 1.2) [15, 16].

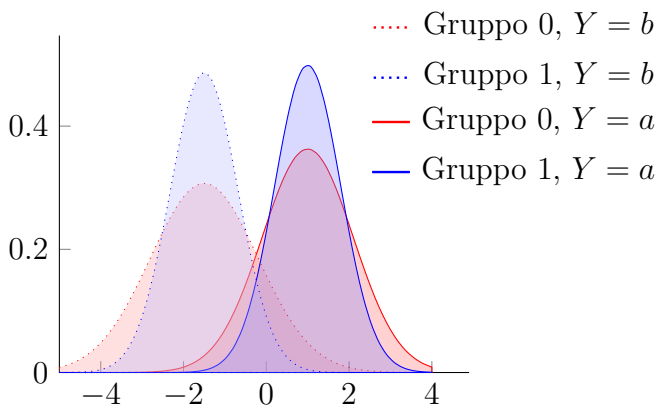

Figura 1.2. **Distribuzione delle predizioni.** L'area in cui le gaussiane dello stesso gruppo si intersecano indicano i FP/FN. La varianza maggiore del gruppo 0 rispetto al gruppo 1 indica una minore accuratezza dell'algoritmo per la data classe.

1.3 Incompatibilità tra le definizioni

Molti lavori dimostrano matematicamente che le definizioni di equità individuale e statistica sono tra loro incompatibili [17], al contrario altri dimostrano che l'incompatibilità è solo apparente e dovuta al numero di semplificazioni e assunzioni internamente contrastanti per rappresentare matematicamente lo spazio d'interesse, che una volta identificate e rimodellate rendono la *fairness* dell'algoritmo individuabile in entrambe le definizioni [18].

La scelta del criterio di equità dipende dal compito che l'algoritmo deve eseguire, dagli spazi con i quali interagisce e dai metodi di modellazione delle caratteristiche che vengono scelti in modo da catturare al meglio le caratteristiche di interesse e conoscere a priori eventuali *trade-off*.

1.3.1 Scelta della definizione in funzione dello spazio

Il lavoro di Friedler et al. [19] identifica due possibili visioni di un determinato spazio: la visione WYSIWYG (What You See Is What You Get) e la visione di

bias strutturale[4]; le due definizioni supportano in modo diverso i criteri di *fairness* discussi fin'ora.

Per la visione WYSIWYG si assume che la mappatura tra lo spazio rappresentato e quello osservato sia priva di distorsione, mentre per la visione di *bias* strutturale si assume che venga inserito del rumore non uniforme tra i gruppi nella rappresentazione.

É documentato che in presenza di *bias* sistemico un approccio tramite equità individuale si rivela iniquo poiché la distorsione delle caratteristiche tra i due gruppi vengono interpretate come non favorevoli per il gruppo che risente del *bias* andando ad acuirlo [19]. Sotto una visione WYSIWYG solo i meccanismi di equità individuale garantiscono scelte non-discriminatorie mentre sotto una visione di *bias* strutturale le uniche misure che consentono di raggiungere l'equità sono quelle di gruppo.

Risulta quindi fondamentale analizzare il caso di applicazione per scegliere al meglio la rappresentazione degli spazi e le metriche con le quali gestire la *fairness* di un algoritmo.

1.4 Problematiche delle metriche di *fairness*

Vengono ora elencati i principali problemi delle metriche precedentemente descritte: tutte le metriche possiedono intrinsecamente un *trade-off* tra precisione e *fairness*, solitamente più precisione si richiede all'algoritmo più sarà *biased* a meno di misure complementari che saranno discusse in (Sezione 4.2) [16].

Inoltre, come documentato in alcuni studi, per 1.2.2 i risultati possono essere intenzionalmente modificati per rimanere *biased* pur raggiungendo le condizioni di *fairness* attraverso la manipolazione delle caratteristiche del gruppo avvantaggiato

[4]Pratiche e atteggiamenti radicati che producono esiti negativi cronici per le popolazioni minoritarie.

o scegliendo intenzionalmente di aggiungere individui di una classe in particolare per il gruppo sfavorito per modificare i parametri di FP/FN [12, 20].

Capitolo 2

ML nei sistemi sanitari

Come illustrato già nell'Introduzione, sistemi di ML vengono utilizzati in ambito sanitario per molteplici scopi, in particolare questo elaborato prende in considerazione il lavoro di Obeymeyer et al. riguardo algoritmi per la gestione delle cure preventive, nel corso del capitolo verrà descritto il suo funzionamento e i risultati ottenuti in ambito di *fairness*. Lo scopo prefissato da questi algoritmi e con i quali sono pubblicizzati è quello di utilizzare i dati provenienti da fascicoli sanitari elettronici (EHR) per garantire a pazienti (con assicurazioni basate sul rischio[1]) ad alto rischio accessi anticipati per evitare visite d'emergenza o ospedalizzazioni, segnalare ai medici pazienti a medio rischio per effettuare visite necessarie ad evitare il peggioramento delle loro condizioni di salute; riducendo in questo modo costi sia per i pazienti che per le aziende ospedaliere [21].

2.1 L'algoritmo

Il codice utilizzato nello studio di Obermeyer et al., fornito sotto licenza open source, simula il comportamento degli algoritmi utilizzati dalle compagnie assicurative in ambito sanitario descritti sopra, il training è realizzato attraverso regressione lasso con penalty calcolata tramite *10-fold cross validation*. L'algoritmo è *race*

[1]Contratti che fanno uso degli algoritmi di valutazione trattati.

blind, cioè non ha informazioni sull'etnia degli individui, segue il codice (Figura 2.1) di come vengono ottenute dal data frame, descritto in (Sezione 2.3), informazioni demografiche escludendo l'etnia:

```
#estrae le caratteristiche demografiche dal data frame
def get_dem_features(df):
    dem_features = []
    prefix = 'dem_'
    for col in df.columns:
        if prefix == col[:len(prefix)]:
            if 'race' not in col:
                dem_features.append(col)
    return dem_features
```

Figura 2.1. **features.py** Se nelle colonne compare *race* il dato viene scartato.

2.2 Il funzionamento

Il rischio finanziario, i costi totali e le risorse economiche consumate dai pazienti sono etichette[2] comuni a tutti i sistemi nonostante la vasta gamma di algoritmi presenti sul mercato statunitense, come documentato dall'*Association for Computing Machinery* [22]. Gli algoritmi restituiscono poi una percentuale di rischio $r \in R$ per ogni paziente che li colloca in 3 diverse zone di rischio; se $r \geq 55\%$ il paziente viene indicato al proprio medico curante insieme a dati aggiuntivi per mitigare successivi peggioramenti mentre se $r \geq 97\%$ viene automaticamente identificato per l'iscrizione al programma, che comporta un'insieme di terapie e visite mediche da effettuare (l'effettiva iscrizione non è garantita, i pazienti devono possedere un contratto assicurativo basato sul rischio e potrebbero non soddisfare altri criteri

[2]Valore che l'algoritmo deve predire per un oggetto in funzione delle caratteristiche dello stesso oggetto.

non meglio precisati). L'algoritmo viene poi lanciato 3 volte all'anno in modo da eseguire un monitoraggio continuo.

2.3 I dati utilizzati

L'algoritmo lavora su un data frame composto dai seguenti elementi [3]:

- *Etichetta costi*: Costi totali per l'anno t, da predire.
- *Vettore caratteristico*: comprende informazioni demografiche (esclusa l'etnia), tipo di assicurazione, diagnosi ICD-9 (classificazione internazionale delle malattie), medicinali prescritti, visite categorizzate (e.g., radiologica, pneumologica, etc.), importi fatturati categorizzati (e.g., dialisi, etc.), tutte riguardo l'anno $t-1$.

In particolare il data frame è stato ottenuto attraverso una collaborazione con un'azienda ospedaliera universitaria non nota [3] e contiene 100.009,00 $\frac{anni}{paziente}$ formati da $\simeq$ 50000 persone, di età media $\simeq$ 50 anni raccolti tra il 2013 e il 2015; composti per il 87,7% da persone bianche e per il 62,8% da individui di sesso femminile. I dati utilizzati comprendono quelli di individui che non possiedono contratti assicurativi basati sul rischio sotto il parere dell'IRB affermando che il consenso dei pazienti non sia necessario poiché l'uso di dati di routine non costituisce rischi.

L'algoritmo restituisce poi l'etichetta $C_{i,t}$ (costi totali del paziente i per l'anno t) e un R_i per l'anno t in funzione di $C_{i,t}$.

2.4 Come sono state effettuate le misurazioni

Lo studio pone come obiettivo la misurazione della *fairness* dell'algoritmo sui gruppi (definiti sull'etnia) attraverso la calibrazione, un postulato dell'equità di gruppo, in sintonia con il lavoro di Friedler et al. secondo i quali rappresentando lo spazio attraverso una visione di *bias* sistemico solo questa misura consente di

raggiungere risultati equi; lo scopo dello studio è ottenere risultati di interesse sociale e storico sulle differenze razziali in ambito di salute tra pazienti che si sono auto-identificati come Neri o Bianchi.

Prendendo in considerazione $H(r)$, misura della salute rispetto al punteggio di rischio [2].

H viene costruita in base al numero di malattie croniche attive, non solo in base alla presenza delle suddette malattie ma tenendo in conto il grado di gestione (terapie attive) delle stesse, e ai valori di marcatori biologici comuni (e.g., colesterolo LDL, glicemia, etc.) con peso in base al loro valore (basso, normale, alto).

Tabella 2.1: **Distribuzione delle comorbidità per classe** [3].

Comorbidità	Bianchi	Neri
Numero di malattie attive	1.26	2.06
Ipertensione (%)	0.32	0.46
Diabete, senza complicazioni	0.08	0.25
Aritmia	0.09	0.10
Ipotiroidismo	0.10	0.06
Obesità	0.08	0.19
Malattie polmonari	0.10	0.14
Cancro	0.09	0.10
Depressione	0.06	0.09
Anemia	0.06	0.11
Artrite	0.04	0.07
Insufficienza renale	0.03	0.08
Squilibri elettrolitici	0.03	0.06
Arresto cardiaco	0.03	0.05
Psicosi	0.03	0.05
Valvulopatia cardiaca	0.03	0.04
Ictus	0.00	0.01
Arteriopatia periferica	0.02	0.04
Diabete, con complicazioni	0.01	0.08
Attacco cardiaco	0.02	0.03
Malattie al fegato	0.01	0.02

2.5 Risultati dello studio

Si nota che prendendo in considerazione $C_{i,t}$ etichetta dei costi (anch'essa predetta dall'algoritmo) l'algoritmo risulta *unbiased*: al percentile 55 i costi per pazienti bianchi sono stati 4995$ e 5146$ (Dollari statunitensi) per pazienti neri, similarmente per i costi nella zona $r \geq 97\%$ soddisfacendo la definizione di calibrazione.

Analizzando i marcatori biologici utilizzati per la funzione H è possibile individuare una differenza tra i due gruppi in funzione di R, al 97esimo percentile (zona dei pazienti a più alto rischio), persone nere hanno ipertensione, diabete, anemia e insufficienza renale più gravi, come replicato in (Figura 2.3), rispetto a persone bianche e 4.8 vs 3.8 malattie in media [2] similarmente a (Tabella 2.1).

Utilizzando la calibrazione (Sezione 1.2.2) che ha una formulazione simile a quella appena riportata è possibile notare come l'algoritmo non soddisfi le condizioni di equità statistica, lo stesso utilizzando la parità demografica: è possibile notare come in $55 \leq r \leq 97$ la percentuale di persone nere non soddisfi l'equazione infatti la percentuale corretta simulata aumenterebbe del 31,8% (Figura 2.2).

Allo stesso modo non viene soddisfatta l'equità individuale, come verificato di seguito:

Prendendo una coppia di individui $\{i, j\}$ con etichetta $C_{i,t} = C_{j,t}$ e $\overrightarrow{X_i} \simeq \overrightarrow{X_j}$ vettori dei bio-marcatori, appartenenti a due gruppi diversi tali che $A_i \neq A_j$; per (Sezione 1.2.1) $\rightarrow R_i \simeq R_j$ a meno di un ϵ invece analizzando il data frame sono individuabili coppie con differenze nel punteggio di rischio discrete, ad esempio (vengono tagliate altre 140 colonne simili):

risk_score_t	program_enrolled_t	cost_t	bps_mean_t	ghba1c_mean_t	hct_mean_t	cre_mean_t	ldl_mean_t	race	dem_female
8,83302E+15	0	9700	118	NA	41.2	1.02	NA	white	0
5,36776E+15	0	9700	123	5.9	44	0.92	NA	black	1

Questa differenza marcata tra le comorbidità (Tabella 2.1) e il punteggio di rischio tra due gruppi si traduce poi in differenze nei trattamenti sanitari: le persone nere hanno spese maggiori in visite d'emergenza e ospedalizzazioni mentre persone

bianche in cure domestiche e visite ambulatoriali [3], rendendo il coordinamento delle cure impari.

Risulta dunque che l'etichetta $C_{i,t}$ non consente di ottenere direttamente le condizioni di *fairness*, nel (Capitolo 4) vengono trattate le scelte di altre etichette che forniscono risultati migliori.

2.5.1 Replicazione dei risultati

Gli autori rendono disponibili un data frame [3] che replica l'originale 48.784,00 $anni*paziente$ vs 100.009,00 $anni*paziente$ e il codice per graficare le funzioni dei bio-marker e delle comorbidità rispetto al punteggio, vengono riportati ora i risultati della funzione H in funzione di R e dei bio-marker in funzione di R comparandoli con quelli dichiarati.

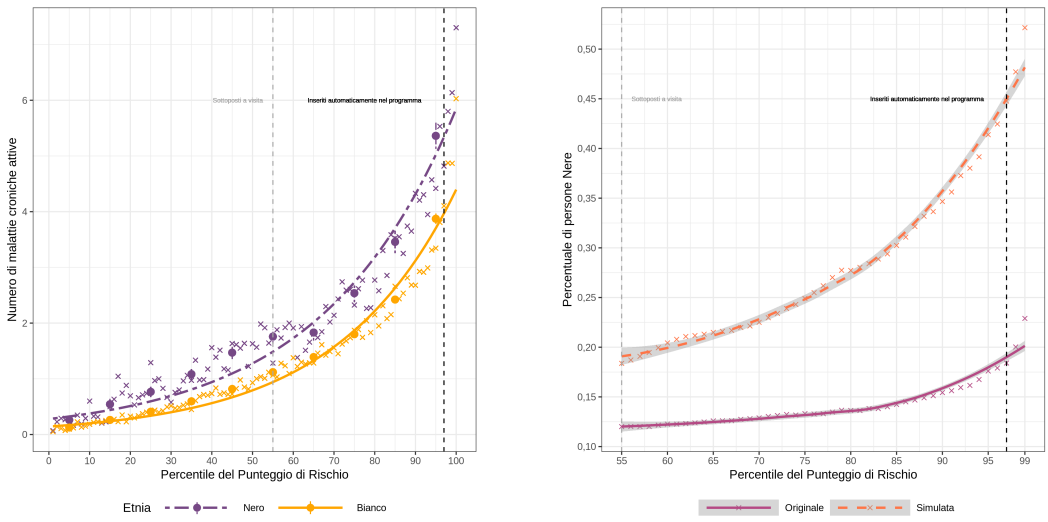

Figura 2.2. **Numero di comorbidità in funzione del punteggi di rischio, per etnia**. (**A**) Confronto delle comorbidità tra i due gruppi a parità di punteggio. (**B**) Confronto della percentuale di persone nere dal 55esimo percentile tra l'algoritmo originale e uno *unbiased*.

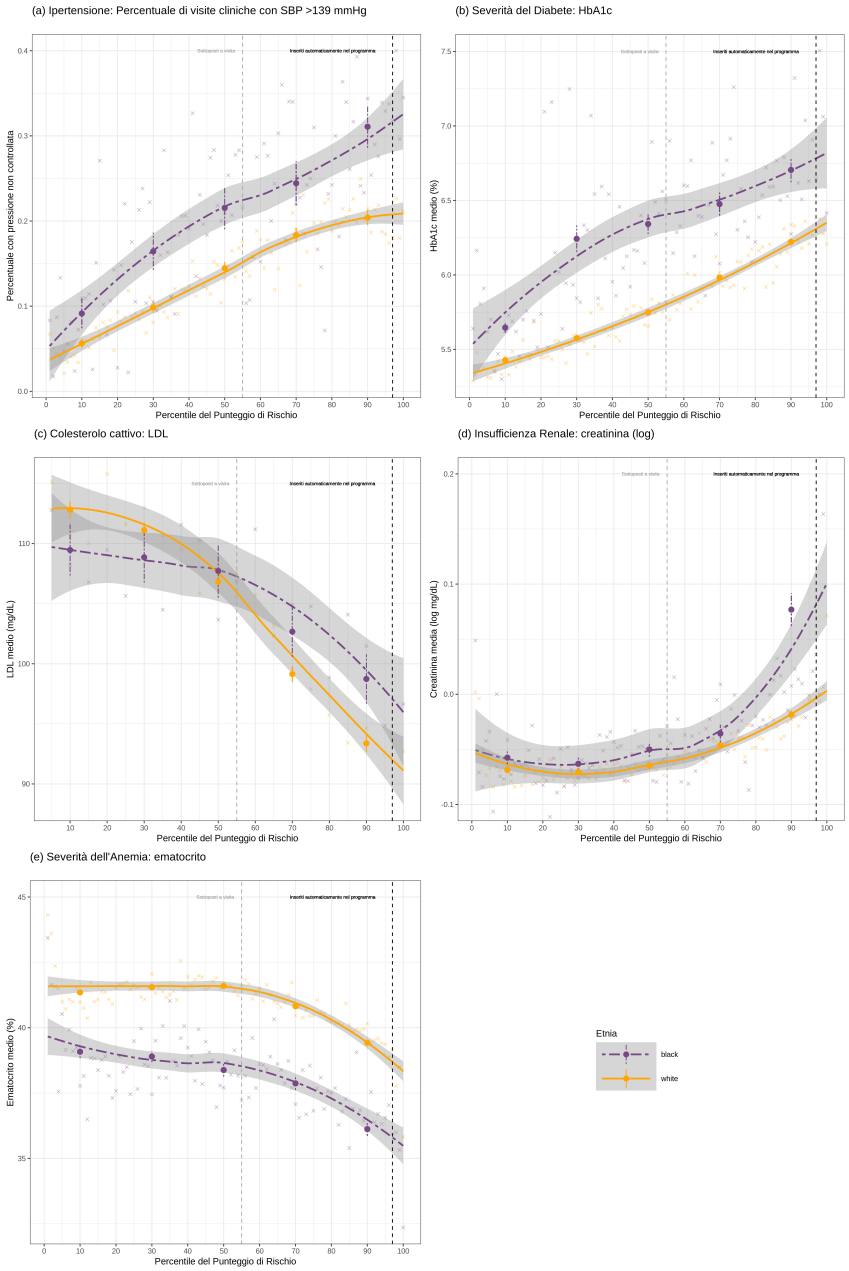

Figura 2.3. **Valori dei bio-marcatori in funzione del punteggio di rischio.** È possibile notare come i valori per le persone nere siano sempre peggiori [23].

I risultati riprodotti in (Figura 2.2) risultano uguali a quelli esposti in [2] alla terza cifra decimale, mentre in (Figura 2.3) le sotto-figure (a), (d), (e) sono sovrapponibili a quelle in [2] mentre (b), (c) seguono un andamento simile.

Capitolo 3

Origine del *bias*

Viene trattata ora l'identificazione delle possibili fonti di *bias* nella raccolta e manipolazione dei dati e nella realizzazione degli algoritmi, con un focus sul data set utilizzato da Obermeyer et al. [3].

3.1 *Bias* nei dati

I dati sanitari utilizzati per il ML derivano da EHR, spazio nel quale successivamente lavoreranno, essi devono risultare completi, di qualità e significativi per l'applicazione. Viene ora mostrato (Figura 3.1) il percorso generale nel quale i dati viaggiano, ogni step è una possibile fonte di *bias*.

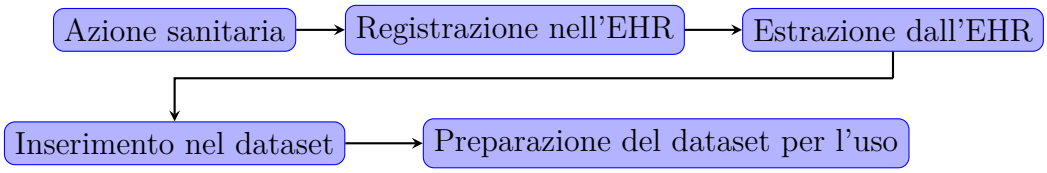

Figura 3.1. **Percorso dei dati**

Eventuali *bias* iniziano a presentarsi ben prima dell'azione sanitaria nei dati a causa del modello reale (eg. come accade nel caso di Obermeyer et al.).

Essendo i dati risultati di decisioni passate, ereditano i *bias* della società: studi dimostrano come persone nere ricevano esiti sbagliati con più probabilità di persone

bianche, risultando in una minore fiducia delle persone nere verso il sistema sanitario che porta a rivolgersi meno spesso ad esso per eventi non critici [34], i medici siano più portati a riportare umore basso che diagnosticare depressione per le donne [27], i medici di base siano più titubanti nel registrare problemi di abuso d'alcol se il dato sarà condiviso con specialisti e servizi di salute mentale [32] e pazienti senza assicurazione sanitaria ricevano assistenza sub-standard con più probabilità rispetto a quelli con assicurazione [24].

Il *bias* può essere intensificato anche dalla presenza di sistemi di rimborso (alcuni lavori rilevano *upcoding*[1] nel sistema sanitario statunitense [25]) e dal fatto che alcuni gruppi siano sistematicamente sotto-testati associandoli successivamente ad un rischio minore per determinate patologie [38].

Si delinea quindi l'impossibilità dell'esistenza di dati *fair* e oggettivi a causa di "rumore" sistematico presente nel modello reale, pertanto, durante lo sviluppo, bisogna adottare misure correttive che tengano conto dei seguenti *bias*.

3.1.1 *Bias* di rappresentazione

I *bias* di rappresentazione si presentano quando i dati di input sono sbilanciati tra le classi (non si verifica nello studio di Obermeyer et al., il 12.3% di popolazione nera rispecchia la popolazione generale negli Stati Uniti al censimento del 2020 [26]), cioè lo spazio di input possiede pochi esempi di un particolare sottospazio, questo causa predizioni meno accurate per la classe minoritaria [28].

Ad esempio in uno studio di un algoritmo per il riconoscimento del Parkinson a partire da biomarker solo il 18,6% degli individui inclusi nello studio erano di sesso femminile, come conseguenza, la sotto-rappresentazione della classe portava a predizioni accurate solo per la classe maggioritaria [29].

[1]Fatturazione medica non dettagliata più costosa dei servizi eseguiti.

Tra gruppi non omogenei, uno sbilanciamento tra le classi causa una più difficile interpretazione dei pattern del gruppo minoritario, allo stesso modo la presenza maggioritaria di dati incompleti per una classe; fino al 2017 solo il 54% degli studi su algoritmi di predizione basati su dati da EHR ha tenuto conto dei dati mancanti [30], integrando data set esterni per includere una popolazione variegata o utilizzando strumenti per creare istanze delle classi sotto-rappresentate.

Bias di rappresentazione di diverso tipo possono presentarsi anche a causa di scelte mediche differenti, ad esempio nel Regno Unito linee guida promuovono la raccolta di dati sulla pressione cardiaca di tutta la popolazione [31] mentre nei Paesi bassi è promossa solo per pazienti che soffrono di malattie croniche [32]. Questo porta ad una differenza nel numero di occasioni di *data-recording* che forniscono conseguentemente dati di platee diverse (l'intera popolazione per il Regno Unito, la popolazione con malattie croniche per i Paesi bassi). Alcuni protocolli invece raccomandano di non inserire in dataset nazionali eventi rari (eg. tassi di mortalità locali bassi [33]); nonostante queste omissioni siano fatte per preservare la privacy dei pazienti contribuiscono ad introdurre un *bias* di rappresentazione.

La presenza di linee guida per promuovere la raccolta di alcuni dati porta ad un aumento delle occasioni di *data-recording*, al contrario un obbligo in alcune situazioni potrebbe aumentare il carico di lavoro e portare a dati di scarsa qualità [32].

3.1.2 *Bias* di misurazione

Il *bias* si presenta quando i dati tra le classi presentano qualità significativamente diverse, quindi durante l'azione sanitaria. Variazioni nella qualità dei dati dei gruppi portano a differenze nell'accuratezza dell'algoritmo per i gruppi, il *bias* di rappresentazione in questo caso è risultato di un *bias* storico: persone con reddito basso hanno probabilità più alte di recarsi presso ospedali comunitari [34], di essere visitati da tirocinanti (qualità dei dati sistematicamente più bassa [24]) e allo stesso

tempo hanno probabilità più alte di vivere in zone in cui le strutture mediche hanno basso livello di digitalizzazione, portando ad un *data-recording* effettuato su sistemi che non comunicano con quelli nazionali, le conseguenze sono frammentazione dei dati (è più comune in individui di classe socio-economica bassa frequentare diversi luoghi per accedere ai servizi sanitari [24]), o *data-recording* non effettuato; allo stesso modo EHR potrebbero contenere solo i casi più gravi per un certo gruppo (eg. solo visite d'emergenza) perché i dati sulla salute non sono raccolti fino a quando non ci si reca in un'azienda sanitaria, risultando in dati mancanti per gruppi che hanno ostacoli nell'accesso ai servizi di salute.

I sistemi EHR devono anche soddisfare dei criteri di accessibilità e facilità di utilizzo, vari lavori mettono in luce la complessità dei sistemi e il loro conseguente errato utilizzo, un lavoro sul sistema nazionale sanitario del Regno Unito osserva che su 352 codici per categorizzare sintomatologie allergiche il 10% viene utilizzato nel 95% dei casi mentre il 21% dei codici non è mai stato utilizzato [35].

3.1.3 *Bias* di valutazione

Il *bias* si presenta tra l'estrazione dei dati e l'inserimento nei dataset quando i dati utilizzati durante l'addestramento o la valutazione delle performance dell'algoritmo non sono significativi per il caso di studio.

Nel caso di Obermeyer et al. questo *bias* non si presenta in quanto la valutazione e il training sono effettuati su dati provenienti dalla stessa sorgente sulla quale l'algoritmo lavorerà.

È importante, quando le sorgenti dei dati risultano diverse, che esse siano compatibili, una grande quantità di dati non sostituisce dati di qualità [32], alcuni sistemi di EHR fanno utilizzo di diversi sistemi di classificazione per le diagnosi e le malattie con complessità diversa (eg. l'International Classification of Primary Care fa uso di circa 600 codici mentre altri sistemi potrebbero utilizzare una versione semplificata causando perdita di informazione nel mapping ad una versione

semplificata o dati mancanti nel mapping inverso) [32].

3.2 *Bias* nei modelli

Il *bias* presente nei dati si riflette nei modelli (Sezione 3.2.2) ma può essere introdotto in questo livello a causa di un implementazione errata (Sezione 3.2.1).

3.2.1 *Bias* di aggregazione

Si verifica successivamente alla preparazione dei dataset per l'uso, quando si assume che un modello uno-per-tutti sia capace di soddisfare le esigenze dei sottogruppi; la correlazione tra le feature può essere diversa tra i vari gruppi (e.g. l'intolleranza al lattosio è significativamente più comune in alcuni gruppi etnici [36]), in questo caso a meno di un fine-tuning sui diversi gruppi i pattern vengono sviluppati su quello dominante e vengono successivamente applicati anche ad altri gruppi sacrificando l'accuratezza per questi ultimi [37].

3.2.2 Etichette *biased*

Un'etichetta è definita *biased* quando è generata da una funzione di labeling che esprime risultati *biased* in presenza di una verità nota a priori *unbiased*. La scelta di un'etichetta deve essere effettuata tenendo conto dei *bias* intrinsechi all'etichetta stessa.

Nonostante l'utilizzo di etichette sui costi per trovare pazienti con più bisogni di cure sia stato consigliato dall'Istituto di Medicina degli Stati Uniti (NAM) [38] il mezzo si rivela imperfetto a causa dei *bias* nei dati discussi sopra. Lo stesso si verifica anche nel caso frequentemente utilizzato nella letteratura dell'approvazione dei prestiti, un algoritmo addestrato per predire l'approvazione di prestiti rispetto a scelte passate fornisce esiti *biased* a causa di *bias* storici rispetto ad un algoritmo addestrato sugli esiti dei prestiti [28].

Nel caso di Obermeyer et al. le scelte sulla salute risultano scelte sui costi (R è calcolato in funzione di C), seppur i costi non mostrino disparità i rischi di salute risultano profondamente *biased*; un paziente a basso costo nei dati potrebbe esserlo perché non può sostenere le spese mediche o avere difficoltà nell'accesso alle cure.

Capitolo 4

Mitigazione del *bias*

Vengono ora esposti i risultati ottenuti da Obermeyer et al. in materia di *de-biasing*, successivamente vengono testati e confrontati con i risultati originali altri due metodi di *de-biasing*, in seguito vengono descritti altri metodi per la mitigazione e la prevenzione del *bias*.

4.1 Mitigazione del *bias* per il dataset

4.1.1 Scelta delle label

Come descritto in (Sezione 2.5) il *bias* è risultato di una scelta incorretta dell'etichetta da utilizzare pertanto Obermeyer et al. utilizzano 3 varianti dell'algoritmo originale per predire etichette diverse e confrontare i risultati:

1. *Costi totali* nell'anno t: l'etichetta descritta in (Sezione 2.3).

2. *Costi evitabili* nell'anno t: a causa di ospedalizzazioni e visite d'urgenza, attraverso misure preventive (etichetta utilizzata anche nel lavoro di Tamang et al. sulla popolazione danese ma sulla quale non effettuano valutazioni in ambito di *fairness* [39]).

3. *Malattie croniche attive* (MCA) nell'anno t: misura delle malattie attive e dalle malattie in fase di acutizzazione (non è una somma di quelle già presenti nel data set, altrimenti non ci sarebbe bisogno di una predizione).

Risultati ottenuti

Gli algoritmi vengono utilizzati per la predizione non solo delle etichette per i quali sono addestrati ma anche le etichette utilizzate dalle altre varianti, i risultati indicano una grande varianza tra le varianti per quanto riguarda la percentuale di pazienti neri nei gruppi di rischio più alti: dal 14,1% della variante dei costi totali al 26,7% della variante MCA [2].

È possibile notare come il predittore addestrato sull'etichetta Costi totali sia quello con le performance peggiori in termini di *fairness*, mentre il predittore addestrato sull'etichetta Costi evitabili invece risulti quello con fasce dal $\geq$ 97esimo percentile più popolose, dal 16,5% di pazienti dell'algoritmo originale al 26,8%.

Complessivamente viene misurata una riduzione del *bias* molto superiore alla riduzione del *bias* compiuta dai medici (si ricordi che l'assegnazione da parte dell'algoritmo alla fascia a rischio è condizione sufficiente ma non necessaria), che possono far accedere al programma (Sezione 2.2) anche pazienti che l'algoritmo ha classificato nella zona $< 97\%$, portando la percentuale di individui con etnia nera che hanno avuto accesso al programma al 19,2%.

I risultati di Obermeyer et al. sono stati condivisi successivamente con il produttore dell'algoritmo utilizzato dall'ospedale il quale ha indipendentemente replicato le analisi, rilevando i *bias* descritti da Obermeyer et al., e ha successivamente stipulato un accordo per tradurre i risultati ottenuti in un futuro algoritmo che utilizzi un indice che unisca le comorbidità attive e future insieme ai costi [2].

Nella (Tabella 4.1) e (Tabella 4.2) vengono replicati i calcoli con le diverse varianti utilizzando il data set fornito da Obermeyer et al. descritto in (Sezione 2.3),

si può notare come i risultati siano identici a quelli riportati da Obermeyer at al. a meno di un $\epsilon = 0{,}001$.

Tabella 4.1. **Confronto tra le etichette Costi totali e Costi evitabili** Etichetta del training (righe) e concentrazione per etichetta (colonne) di individui al o sopra al 97esimo percentile. Le colonne con (SE) indicano l'errore standard.

Predittore	Costi totali	Costi totali SE	Costi evitabili	Costi evitabili SE
Costi	0,165	0,003	0,234	0,003
Costi evi	0,157	0,003	0,268	0,003
Malattie croniche	0,142	0,003	0,245	0,003
Differenza migliore-pe	0,023		0,034	

Tabella 4.2. **Etichetta MCA e percentuale di persone nere**

Predittore	Malattie croniche attive	Malattie croniche attive SE	Etnia nera	Etnia nera SE
Costi totali	0,121	0,002	0,192	0,003
Costi evitabili	0,152	0,003	0,259	0,003
Malattie croniche attive	0,165	0,003	0,284	0,003
Differenza migliore-peggiore	0,044		0,092	

4.1.2 Ulteriori test

Vengono proposti ora due test sulla *fairness* utilizzando una *label* diversa e un *pre-processing* dei dati.

Training con *label* Punteggio di Rischio

Questo test viene eseguito sulla falsariga dei test effettuati in [3], utilizzando in questo caso come *label* per il training il punteggio di rischio R_i (Figura 4.1); come descritto in (Sezione 2.1) l'algoritmo base ottiene il punteggio di rischio derivandolo dall'etichetta $C_{i,t}$ (*Costi totali*); i punteggi sul quale l'algoritmo è addestrato sono quelli calcolati in seguito all'esecuzione dell'algoritmo base.

```
# Y da etichettare
    Y_predictors = ['risk_score_t']

#training con label risk score
    risk_score_r2_df, \
    pred_risk_score_df, \
    risk_score_lasso_coef_df = model.train_lasso(train_df,
                                        holdout_df,
                                        x_column_names,
                                        y_col='risk_score_t',
                                        outcomes=Y_predictors,
                                        n_folds=n_folds,
                                        include_race=
                                          include_race,
                                        plot=save_plot,
                                        output_dir=OUTPUT_DIR)
```

Figura 4.1. **main.py**

I dati mostrano un risultato simile all'algoritmo originale rispetto la concentrazione di individui nella fascia $\geq 97\%$ mentre mostrano un risultato intermedio riguardo la percentuale di persone nere e le malattie attive in questa fascia rispetto all'algoritmo base e quelli addestrati sulle etichette *Costi evitabili* e *Malattie croniche attive*.

Tabella 4.3. **Risultati con etichetta Punteggio di Rischio** ottenuti modificando **table2.py**

Predittore	Punteggio di Rischio	Punteggio di Rischio SE	Malattie croniche attive	Malattie croniche attive SE	Etnia nera	Etnia nera SE
Punteggio di Rischio	0,167	0,003	0,146	0,003	0,238	0,003

Pre-processing

Il seguente test consiste nell'utilizzare una metodologia discussa in (Sezione 4.2.1), attraverso l'uso della libreria *fairlearn* (Figura 4.2) [40].

Il metodo consiste nell'applicazione di una trasformazione lineare alle *features* per eliminare la correlazione tra queste e quelle definite come "sensibili", in seguito a varie combinazioni i risultati migliori sono stati ottenuti inserendo tra le caratteristiche sensibili i seguenti dati:

- *dem_race_black:* indicatore binario sull'etnia dell'individuo;
- *cost_emergency_tm1:* costi per visite d'emergenza nell'anno $t-1$;
- *gagne_sum_tm1:* somma delle comorbidità per l'anno $t-1$.

I risultati, mostrati in (Tabella 4.4) e (Tabella 4.5), riportano un aumento per gli algoritmi addestrati sulle etichette *Costi evitabili* e *Malattie croniche attive* della percentuale di individui di etnia nera nella fascia $\geq 97\%$, mentre risultati peggiori riguardo l'etichetta *Costi totali*; le percentuali di individui nella fascia $\geq 97\%$ rimangono invariate.

```
from fairlearn.preprocessing import CorrelationRemover
#lista di feature sensibili
list_sen = ['dem_race_black',
            'cost_emergency_tm1', 'gagne_sum_tm1']
#creiamo un oggetto decorrelatore
cor = CorrelationRemover(sensitive_feature_ids=list_sen)
#fit del data frame e trasformazione
cor.fit_transform(train_X, train_y)
```

Figura 4.2. **model.py**

Tabella 4.4. **Confronto tra le etichette Costi totali e Costi evitabili**
Etichetta del training (righe) e concentrazione per etichetta (colonne) di individui al o sopra al 97esimo percentile. Le colonne con (SE) indicano l'errore standard.

Predittore	Costi totali	Costi totali SE	Costi evitabili	Costi evitabili SE
Costi totali	0,170	0,003	0,241	0,003
Costi evitabili	0,157	0,003	0,268	0,003
Malattie croniche attive	0,143	0,003	0,245	0,003
Differenza migliore-peggiore	0,027		0,027	

Tabella 4.5. **Etichetta MCA e percentuale di persone nere**

Predittore	Malattie croniche attive	Malattie croniche attive SE	Etnia nera	Etnia nera SE
Costi totali	0,121	0,003	0,164	0,003
Costi evitabili	0,152	0,003	0,288	0,003
Malattie croniche attive	0,165	0,003	0,288	0,003
Differenza migliore-peggiore	0,044		0,124	

4.2 Altri metodi di mitigazione

4.2.1 *Pre-processing* dei dataset

Questo approccio ha come scopo il rimuovere il *bias* attraverso la trasformazione dei dati (attraverso metodi di *whitening* o *re-weighing*) pur mantenendo il massimo livello di informazione ricavata dagli stessi. La maggior parte degli approcci utilizza spazi in cui le *features* protette sono ortogonali allo spazio stesso il che impedisce all'algoritmo di dedurre informazioni sulle *features* protette oppure assegnando alle diverse tuple nel *dataset* dei pesi (quindi senza manipolare i dati originali), scegliendo correttamente i pesi da calcolare sulle caratteristiche protette si riescono ad ottenere dei risultati *unbiased* [28, 41].

Altri approcci consistono nell'utilizzo di *adversarial networks* per modellare le *features* protette; massimizzando l'accuratezza del predittore e minimizzando la capacità della rete avversaria di predire gli attributi protetti [28], utilizzata con successo in ambito medico per rilevare lo stato di COVID-19 in individui ammessi in ospedale in modo *unbiased* tra diversi gruppi demografici [42].

4.2.2 *Post-processing* dei risultati

L'approccio risulta particolarmente utile quando i dati o l'algoritmo non sono disponibili, nella maggior parte dei casi vengono modificati i parametri in uscita per la classe *biased* in modo da soddisfare i requisiti per le metriche di equità di gruppo

mentre requisiti per equità individuale risultano molto difficili da raggiungere senza sacrificare le performance con questo metodo [28, 43, 44].

Obermeyer et al. effettuano un test per stimare la percentuale di individui con etnia nera che dovrebbero far parte della zona ad alto rischio (misurazione del *bias* del modello): scambiando individui tra la zona $\geq 97\%$ e $< 97\%$ in base al numero di comorbidità fino a quando gli individui sul margine possiedono lo stesso numero medio di comorbidità, come mostrato in (Sezione 2.5.1 - Figura 2.2), ottenendo un percentuale del 46,5% di individui con etnia nera rispetto al 17,7% ottenuto dal modello base (percentuale superiore anche a quelle delle varianti che soddisfano le condizioni di *fairness*) [2]. Il metodo, attraverso un fine tuning per il soddisfacimento della calibrazione (Sezione 1.2.2), potrebbe essere utilizzato come post-processer per il problema stesso.

4.2.3 Feature augmentation

I metodi (eg. SMOTE, Fair-SMOTE), che fanno parte della famiglia dell'*oversampling*, consistono nel generare pseudo-istanze nelle vicinanze dei gruppi minoritari cercando istanze dei suddetti gruppi e i loro k-*nearest neighbors* per aggiungere al data set pseudo-istanze intermedie a questi, in modo da eliminare *bias* di rappresentazione; variazioni sull'algoritmo come Fair-SMOTE vanno a bilanciare anche le frequenze degli attributi sensibili andando ad aumentare il valore di recall (Sezione 1.2.2), vari lavori evidenziano la mancanza di effetti collaterali dei metodi sulle performance generali dei gruppi non scelti come target [28, 45, 46] e la loro efficacia in ambito di algoritmi per la salute [47] (eg. aumentando la precisione nella classificazione di gonartrosi in soggetti di età elevata o obesi [48]).

4.3 Prevenzione del *bias* e linee guida

Altrettanti lavori sottolineano l'importanza della creazione e l'utilizzo di *framework* per ridurre al minimo a priori il rischio di *bias* nei dati e nei risultati [34, 49, 50]. Vengono riportati ora alcuni dei lavori che hanno ottenuto risultati notevoli.

4.3.1 Interpretabilità del sistema

L'interpretabilità deve essere ricercata e inclusa a partire dalla concezione del sistema: esso deve risultare familiare ai processi cognitivi umani applicati dai medici e supportato dall'evidenza scientifica per evitare di perpetuare pratiche cliniche non ottimali; il tutto deve essere affiancato da "algoritmovigilanza" [34, 47], una rivalutazione sui benefici apportati attraverso le diverse popolazioni (eg. etnia, sesso, età), soprattutto nei sistemi con apprendimento naturale (dove l'addestramento continua anche in *working mode*), a causa di una possibile traslazione delle distribuzioni nel tempo che porta ad un declino delle prestazioni [51].

4.3.2 Ottenimento degli attributi protetti

Molti stati possiedono regole stringenti sulla raccolta di dati sensibili come l'etnia vietandola o lasciando la possibilità di auto-identificazione agli individui; in alcuni casi la presenza di attributi sull'etnia è necessaria per utilizzare i sub-set più corretti (eg. analisi di biomarcatori, analisi sulla *fairness*, analisi della qualità dei dati per gruppo), uno dei risultati ottenuti in (Sezione 2.1) ha mostrato come l'assenza di questi dati non sia realmente rilevante per la *fairness* durante il training (i risultati dell'algoritmo che fa uso del dato sull'etnia e quello che lo rimuove sono sovrapponibili) ma fondamentale per alcuni metodi di *de-biasing*; in tal caso metodi come Bayesian Improved Surname Geocoding Method (consigliati anche dalla National Academy of Medicine) possono essere utilizzati per cercare di ricostruire

il dato protetto non registrato. Il metodo è capace di assegnare ad ogni individuo una percentuale per ognuna delle sei etnie disponibili utilizzando dati geografici, il cognome e un database del censo statunitense; esso ha dimostrato di aver alti livelli di accuratezza per popolazioni più anziane (poiché mostrano una minore mobilità immobiliare) mentre l'accuratezza cala significativamente in caso di individui con etnia mista [52].

Si rimarca quindi l'importanza sia da parte dei legislatori che da parte dei fornitori di servizi per la salute di riuscire a raccogliere e gestire in modo chiaro e sicuro i dati sensibili riguardo l'etnia degli individui per poter far fronte alla gestione della *fairness* degli algoritmi.

4.3.3 Standard per i dati minimi obbligatori

Una parte dei lavori si è concentrata invece su standard (eg. MINIMAR, TRIPOD-ML) per i dati minimi da raccogliere e rendere disponibili agli *stakeholder* allo scopo di capire le popolazioni trattate e i possibili *bias* nei dati, lo scopo è incentivare algoritmi che hanno implementato soluzione in ambito di *fairness*; i dati resi disponibili includono: ragioni di inclusione/esclusione, distribuzione delle etnie e degli status socio-economici, *features* usate, dati da predire, come viene trattata la mancanza dei dati, etc. [53].

4.3.4 Strumenti per la valutazione del *bias*

Altri strumenti che hanno dimostrato buoni risultati sono strumenti di valutazione, simili in alcune parti agli standard per i dati di cui discusso sopra, che permettono una revisione sistematica e di stimare il rischio di *bias* (eg. PROBAST [54]). I seguenti strumenti rendono successivamente visibile a tutti gli *stakeholder* i risultati in termini di *fairness*:

- *Descrivere l'uso finale del modello e lo spazio di lavoro.*

- *Classificare il tipo di predizione, di validazione, i modelli e gli output di interesse.*
- *Classificare i rischi di applicazione e i rischi di bias*: eg. le fonti dei dati e la loro composizione, i criteri di applicabilità.
- *Classificare i domini dei dati*: eg. se i dati sono disponibili per tutte le istanze, se i dati raccolti possiedono la stessa accuratezza per tutte le istanze.

Anche in questo caso lo scopo è incentivare i produttori di algoritmi ad affrontare tematiche di *fairness*.

Conclusioni

L'elaborato ha affrontato le tematiche di *fairness* in ambito di algoritmi per la salute attraverso uno dei maggiori esempi in letteratura; si è delineata l'importanza dell'analisi degli spazi di lavoro, la scelta non triviale delle metriche da utilizzare per assicurare la *fairness* dell'algoritmo e la stesura di documentazioni che possano rendere visibile agli *stakeholder* le scelte fatte e i risultati ottenuti.

Di altrettanto valore la trasparenza dei dati, la possibilità di utilizzare *framework* comuni per la compatibilità di dati da diverse fonti che potrebbero raccogliere diverse comunità minoritarie per migliorare i dati di training stessi e l'importanza dell'inclusione nello sviluppo di risorse riguardanti non solo i gruppi maggioritari per il caso di studio.

I risultati ottenuti in letteratura sono notevoli, tenendo conto dell'inesistenza di metodi generali applicabili a tutte le situazioni, le soluzioni su misura (eg. di Baker et al. [11], Friedler et al. [19] per algoritmi di valutazione della qualità dell'insegnamento, di Fong et al. [16], Mishler et al. [43], Moustakidis et al. [48] per algoritmi in ambito sanitario, etc.) trattano e pongono delle basi di analisi per la *fairness* di algoritmi che trovano utilizzo in ambienti simili.

Gli attori principali, legislatori e produttori di algoritmi, devono tenere conto della crescente ricerca, dei costi sociali che l'adozione diffusa di algoritmi sanitari *biased* potrebbe comportare per comunità minoritarie e dell'ondata invernale che potrebbe ricadere sul *machine learning* in caso questo accadesse.

Bibliografia

[1] WIPO, *WIPO Technology Trends 2019: Artificial Intelligence*, Geneva: World Intellectual Property Organization, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf, pp. 39-40, 2019.

[2] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., *Dissecting racial bias in an algorithm used to manage the health of populations*, Science 366, https://escholarship.org/content/qt6h92v832/qt6h92v832.pdf, pp. 447-453, 2019.

[3] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., *Supplementary Materials for Dissecting racial bias in an algorithm used to manage the health of populations*, Science 366, https://escholarship.org/content/qt6h92v832/qt6h92v832.pdf, pp. 447-453, 2019.

[4] Fatima, N., *AI in Photography: Scrutinizing Implementation of Super-Resolution Techniques in Photo-Editors*, 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), https://sci-hub.st/10.1109/ivcnz51579.2020.9290737, pp. 1-6, 2020.

[5] Nabin, K. Mishra, M. Emre, Celebi, *An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning*, https://arxiv.org/abs/1601.07843, 2016.

[6] Carbune, Victor, Gonnet, Pedro, Deselaers, Thomas, Rowley, Henry, Daryin, Alexander, Calvo, Marcos, Wang, Li-Lun, Keysers, Daniel, Feuz, Sandro, Gervais, Philippe, *Fast multi-language LSTM-based online handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR)*, 23, 10.1007/s10032-020-00350-4, https://arxiv.org/pdf/1902.10525.pdf, 2020.

[7] Narayanan, A., *21 fairness definitions and their politics*, Technical report, Conference on Fairness, Accountability and Transparency 2018, 2018.

[8] Noble, S., Florida State University Libraries Research Guides, 2016.

[9] Bhardwaj, R., Nambiar, A. R., Dutta, D., *A Study of Machine Learning in Healthcare*, 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8029924, 2017.

[10] Angwin, J., Larson, J., Mattu, S., Kirchner, L., *Machine bias*, ProPublica, https://www.propublica.org/article/

how-we-analyzed-the-compas-recidivism-algorithm, May, 23, 2016.

[11] Baker, R.S., Hawn, A., *Algorithmic Bias in Education*, Int J Artif Intell Educ, https://doi.org/10.1007/s40593-021-00285-9, 2021.

[12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., *Fairness through awareness*, in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214-226. ACM, 2012.

[13] Dwork, C., Ilvento, C., *Individual fairness under composition*, FATML, https://arxiv.org/abs/1806.06122, 2018.

[14] Tolan, S., *Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges*, JRC Working Papers on Digital Economy 2018-10, Joint Research Centre (Seville site), 2018.

[15] Mayson, Sandra G., *Bias In, Bias Out*, https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=1010&context=aidr_collection, 2018.

[16] Fong, H., Kumar, V., Mehrotra, A., Vishnoi, N. K., *Fairness for AUC via Feature Augmentation*, arXiv e-prints, https://arxiv.org/pdf/2111.12823.pdf , 2021.

[17] Kleinberg, J. M., Mullainathan, S., Raghavan, M., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, CoRR, abs/1609.05807, http://arxiv.org/abs/1609.05807, 2016.

[18] Reuben Binns, *On the Apparent Conflict Between Individual and Group Fairness*, CoRR, abs/1912.06883, http://arxiv.org/abs/1912.06883, 2019.

[19] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., *On the (im)possibility of fairness*, CoRR, abs/1609.07236, http://arxiv.org/abs/1609.07236, 2016.

[20] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., *Algorithmic decision making and the cost of fairness*, CoRR, abs/1701.08230, http://arxiv.org/abs/1701.08230, 2017.

[21] Johns Hopkins University, *Johns Hopkins ACG Overview*, 2021.

[22] M. Kay, C. Matuszek, S. A. Munson, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819-3828, https://dl.acm.org/doi/proceedings/10.1145/2702123, 2015.

[23] Padilla, O., Abadie, J., *Blood Tests: Normal Values*, Texas Tech Health Science Center, 2021.

[24] Gianfrancesco, M. A., Tamang, S., Yazdany, J., Schmajuk, G., *Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data*, JAMA Intern Med. 2018;178(11):1544-1547, https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2697394, 2018.

[25] Steinbusch, P. J., Oostenbrink, J. B., Zuurbier, J. J., Schaepkens, F. J., *The risk of upcoding in casemix systems: a comparative study*, Health policy (Amsterdam, Netherlands), 81(2-3), 289-299, https://doi.org/10.1016/j.healthpol.2006.06.002, 2007.

[26] Jones, N., Marks, R., Ramirez, R., Merarys, R. V., *2020 Census Illuminates Racial and Ethnic Composition of the Country*, United States Census Bureau, https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-\population-much-more-multiracial.html#:~:text=In%202020%2C%20the%20Black%20or,million%20and%2012.6%25%20in%202010., 2021.

[27] Fassaert, T., Nielen, M., Verheij, R., Verhoeff, A., Dekker, J., Beekman, A., et al., *Quality of care for anxiety and depression in different ethnic groups by family practitioners in urban areas in the Netherlands*, Gen Hosp Psychiatry 2010, 32(4):368-376, https://pubmed.ncbi.nlm.nih.gov/20633740/, 2010.

[28] Fu, R., Huang, Y., Singh, P. V., *AI and Algorithmic Bias: Source, Detection, Mitigation and Implications*, http://dx.doi.org/10.2139/ssrn.3681517, 2020.

[29] Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Mavridis, N., *Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare*, NPJ digital medicine, 3(1), 1-11, https://www.nature.com/articles/s41746-020-0288-5, 2020.

[30] Goldstein, B. A., Navar, A. M., Pencina, M. J., Ioannidis, J. P., *Opportunities and challenges in developing risk prediction models with electronic health records data*, J Am Med Inform Assoc. ;24(1):198-208, doi:10.1093/jamia/ocw042, 2017.

[31] *Hypertension in adults: diagnosis and management*, NICE Guideline, NG136, https://www.nice.org.uk/guidance/ng136/chapter/recommendations, 2022.

[32] Verheij, R. A., Curcin, V., Delaney, B. C., McGilchrist, M. M., *Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse*, Journal of medical Internet research, 20(5), e185, https://doi.org/10.2196/jmir.9134, 2018.

[33] Tiwari, C., Beyer, K., Rushton, G., *The impact of data suppression on local mortality rates: the case of CDC WONDER*, American journal of public health, 104(8), 1386-1388, https://doi.org/10.2105/AJPH.2014.301900, 2014.

[34] Gervasi, Stephanie S., Chen, Irene Y., Smith-McLallen, Aaron, Sontag, David, Obermeyer, Ziad, Vennera, Michael, Chawla, Ravi, *The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It*, Health Affairs, volume 41, 2, pp. 212-218, https://doi.org/10.1377/hlthaff.2021.01287, 2022.

[35] Mukherjee, M., Wyatt, J. C., Simpson, C. R., Sheikh, A., *Usage of allergy codes in primary care electronic health records: a national evaluation in Scotland*, Allergy, 71(11), 1594-1602. https://doi.org/10.1111/all.12928, 2016.

[36] Simoons, F. J., *Primary adult lactose intolerance and the milking habit: A problem in biological and cultural interrelations*, The American Journal of Digestive Diseases, 14(12), pp. 819-836,https://link.springer.com/article/

10.1007/BF01072224, 1969.

[37] Suresh, Harini, John, V. Guttag, *A framework for understanding unintended consequences of machine learning*, arXiv preprint arXiv:1901.10002 2, https://courses.cs.duke.edu/spring20/compsci342/netid/readings/suresh-guttag-framework.pdf, 2019.

[38] Wiens, J., Price, W. N., Sjoding, M. W., *Diagnosing bias in data-driven algorithms for healthcare* Nat Med 26, pp. 25-26, https://doi.org/10.1038/s41591-019-0726-6, 2020.

[39] Tamang, S., Milstein, A., Sørensen, H. T., et al., *Predicting patient cost blooms in Denmark: a longitudinal population-based study*, BMJ Open 2017;7:e011580, https://bmjopen.bmj.com/content/7/1/e011580, 2017.

[40] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K., *Fairlearn: A toolkit for assessing and improving fairness in AI*, Techreport MSR-TR-2020-32, Microsoft, https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/, 2020.

[41] Kamiran, F., Calders, T., *Data preprocessing techniques for classification without discrimination*, Knowl Inf Syst 33, pp. 1-33, https://doi.org/10.1007/s10115-011-0463-8, 2012.

[42] Yang, J., Soltan, A. A., Yang, Y., Clifton, D. A., *Algorithmic Fairness and Bias Mitigation for Clinical Machine Learning: Insights from Rapid COVID-19 Diagnosis by Adversarial Learning*, medRxiv, non *peer-reviewed* al 05/22, 2022.

[43] Mishler, A., Kennedy, E. H., Chouldechova, A., *Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds*, https://arxiv.org/abs/2009.02841, 2020.

[44] Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., Puri, R., *Bias mitigation post-processing for individual and group fairness*, Icassp 2019, pp. 2847-2851, https://arxiv.org/abs/1812.06135, 2019.

[45] Iosifidis, V., Ntoutsi, E., *Dealing with bias via data augmentation in supervised learning scenarios*, Jo Bates Paul D. Clough Robert Jäschke, 24, http://ceur-ws.org/Vol-2103/paper_5.pdf, 2018.

[46] Chakraborty, J., Majumder, S., Menzies, T., *Bias in machine learning software: why? how? what to do?*, Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 429-440, https://dl.acm.org/doi/pdf/10.1145/3468264.3468537, 2021.

[47] Fletcher, R. R., Nakeshimana, A., Olubeko, O., *Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health*, Frontiers in Artificial Intelligence, 3, 116, https://www.frontiersin.org/articles/10.3389/frai.2020.561802/full, 2021.

[48] Moustakidis, S., Papandrianos, N. I., Christodolou, E., Papageorgiou, E., Tsaopoulos, D., *Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness*, Neural Computing and Applications, pp. 1-13, https://link.springer.com/article/10.1007/s00521-020-05459-5, 2021.

[49] Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., Rus, D., *Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure*, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Association for Computing Machinery, New York, NY, USA, pp. 289-295, https://doi.org/10.1145/3306618.3314243, 2019.

[50] Lee, N. T., Resnick, P., Barton, G., *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*, Brookings Report, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-\best-practices-and-policies-to-reduce-consumer-harms/, 2019.

[51] Char, D. S., Abràmoff, M. D., Feudtner, C., *Identifying Ethical Considerations for Machine Learning Healthcare Applications*, The American journal of bioethics : AJOB, 20(11), pp. 7-17, https://doi.org/10.1080/15265161.2020.1819469, 2020.

[52] Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., Omer, S. B., *Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study*, Health services research, 49(1), pp. 268-283, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3922477/?report=classic 2014.

[53] Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., Shah, N. H., *MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care*, Journal of the American Medical Informatics Association, Volume 27, 12, pp. 2011-2015, https://doi.org/10.1093/jamia/ocaa088, 2020.

[54] Wolff, R. F., Moons, K., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., PROBAST Group, *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies*, Annals of internal medicine, 170(1), pp. 51-58, https://doi.org/10.7326/M18-1376, 2019.