Arianna Ricci
10536266
31/08/2020

STRUMENTI DI PREDIZIONE DEL RISCHIO IN CAMPO GIURIDICO: ANALISI DELLE CRITICITA' LEGATE ALLA LORO EQUITA'

Abstract

Il problema dell'equità degli algoritmi utilizzati in campo giuridico per valutare il fattore di rischio di recidiva di un imputato è stato oggetto di recenti studi. In particolare, voglio qui analizzare se e come sia possibile una forma di discriminazione o iniquità e quali soluzioni sono state proposte, per arrivare a valutare la validità dell'utilizzo di tali algoritmi in base al livello di equità che possono garantire.

La mia tesi è che allo stato attuale il livello di equità di tali strumenti non sia sufficiente a garantirne l'affidabilità e limiti ancora molto il nostro senso di fiducia. Dovremmo quindi cercare di limitarne l'utilizzo finché questo aspetto non verrà significativamente migliorato.

Introduzione

Da quando si è introdotto l'uso degli algoritmi come supporto al nostro processo decisionale, sono state condotte molte indagini sulle problematiche che il loro utilizzo comporta.

Sono state definite quattro caratteristiche fondamentali che tali algoritmi dovrebbero possedere affinché possa esserci un certo grado di fiducia nell'utilizzarli: precisione della predizione, equità e uguaglianza davanti alla legge, trasparenza ed affidabilità ed infine privacy delle informazioni e libertà di espressione (Scantamburlo T., 2018).

La mia indagine si è focalizzata sulla terza caratteristica: equità e uguaglianza di fronte alla legge, nel particolare ambito di una possibile discriminazione razziale che gli strumenti di valutazione del rischio di recidiva di un imputato possono portare in ambito giudiziario.

In particolare ho analizzato le possibili cause di ingiustizia e le possibili soluzioni, partendo da un esempio concreto considerato non equo, concludendo che il livello di equità proposto da tali strumenti di valutazione è ancora insufficiente a garantire una piena fiducia in essi e una completa accettazione del loro utilizzo come supporto alla decisione umana.

Ho scelto in particolare questo ambito, quello giudiziario, sia perché nella definizione stessa è indicata "uguaglianza di fronte alla legge" sia perché ritengo sia il più sensibile a questa problematica. Introdurre un qualsiasi tipo di discriminazione o iniquità nel processo decisionale che dovrebbe garantire "la giustizia" è a parer mio un pericoloso paradosso.

Prima di continuare, mi sembra importante esplicitare una premessa: tutta la mia analisi ha come sfondo la realtà statunitanse, poiché è il paese in cui gli strumenti di predizione del rischio vengono maggiormente utilizzati e di conseguenza la maggior parte degli studi condotti fanno riferimento al contesto sociale e giudiziario nord-americano.

Sezione 1: CASO COMPAS

COMPAS ("Correctional Offender Management Profiling for Alternative Sanctions"), sviluppato dall'azienda Northpointe, è uno strumento di valutazione adottato da alcuni stati americani (New York, Wisconsin, California, Florida ed altri) per valutare il rischio di recidiva di un imputato, ossia la probabilità che torni a delinquere, classificando il rischio come basso, medio o alto.

È stato pensato come supporto decisionale per i giudici, in particolare riguardo alla possibilità di concedere la libertà vigilata od altri trattamenti ad un imputato, ma è possibile che venga utilizzato anche in altre fasi processuali e quindi anche per emettere le sentenze.

Tale strumento è diventato oggetto di un grande dibattito in seguito alla pubblicazione da parte della redazione di ProPublica di uno studio che dimostra la presenza di un bias razziale al suo interno. (Angwin, Machine Bias, 2016)

In particolare, lo studio rileva una disparità di punteggio tra gli imputati bianchi e quelli neri, che potrebbe sembrare non significativa poiché vi è effettivamente una maggiore probabilità di recidiva tra gli imputati neri rispetto a quelli bianchi, ma soprattutto rileva una disparità nel tasso di errore della predizione. (Angwin, How We Analyzed the COMPAS Recidivism Algorithm, 2016)

Come riportato nella tabella seguente infatti i neri hanno il doppio della probabilità rispetto ai bianchi di essere etichettati ad alto rischio e non esserlo mentre i bianchi hanno il doppio della probabilità dei neri di essere etichettati a basso rischio e non esserlo.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure 1

In risposta a questo studio (Flores A. W., 2016) hanno criticato sia la raccolta dei dati da parte di ProPublica[1] sia l'analisi condotta e le conseguenti conclusioni raggiunte.

Essi sostengono infatti che la precisione della previsione non varia particolarmente tra i due gruppi di popolazione (69% per gli imputati bianchi e 67% per gli imputati neri) e soprattutto che è naturale che gli imputati neri ricevano punteggi mediamente più alti rispetto a quelli bianchi poiché hanno una più alta probabilità di recidiva.

Il caso COMPAS è un chiaro esempio di come sia possibile valutare la validità di un algoritmo in base a parametri differenti: secondo i creatori dello strumento, infatti, una buona precisione generale (68%) e un simile grado di precisione tra i due diversi gruppi (69% per gli imputati bianchi e 67% per gli imputati neri) è sufficiente a garantire che lo strumento sia valido e privo di bias razziali e discriminatori.

Secondo gli autori di ProPublica, invece, tale discriminazione si insinua nel tasso di errore che differisce in maniera importante e discriminante tra i due gruppi, come si vede in Figura 1.

Nello specifico questa differenza può tradursi in un trattamento differente di questi due gruppi di fronte alla legge: gli imputati bianchi hanno più probabilità di ricevere un punteggio più basso, anche se errato, e quindi avere maggior accesso a misure quali la libertà vigilata o cauzioni più basse, mentre gli imputati neri hanno più probabilità di ricevere un punteggio erroneamente elevato e quindi vedersi negate queste possibilità.[2]

[1] Alcune critiche mosse alla raccolta di dati e alla scelta dei p-valore dei test sono potenzialmente corrette ma comunque la generalizzazione del problema che vado ad affrontare esula da questo particolare studio e i problemi di equità vengono spesso presentati in termini generali o con dati fittizi, quindi corretti.

[2] Un altro importante problema è l'utilizzo di questo strumento in tutte le fasi processuali, che può comportare quindi non solo un diverso accesso alla libertà vigilata o alle misure alternative, ma una vera e propria sproporzione tra la gravità delle sentenze emesse verso gli imputati neri rispetto a quelle emesse verso gli imputati bianchi. Tuttavia, in

Dall'analisi di questi risultati sembra che si possa considerare che l'algoritmo utilizzato da COMPAS sia effettivamente affetto da un bias razziale e possa essere quindi considerato discriminatorio.

Sezione 2: PROBLEMATICHE GENERALI

Da questo caso è possibile evincere dei concetti più generali e delle definizioni più formali dei problemi sollevati.

Infatti, in seguito a questo caso sono stati condotti diversi studi sia per valutare se effettivamente gli algoritmi potessero essere discriminatori sia più in generale per valutarne l'equità. Molta di questa letteratura si focalizza principalmente sugli algoritmi di machine learning utilizzati negli strumenti di predizione del rischio di recidiva negli imputati e prende spesso in considerazione la distinzione di razza tra gli imputati[3].

Considero quindi che abbia senso ampliare l'indagine sui bias razziali ad un più generico studio sull'equità di un algoritmo rispetto a due diversi gruppi di popolazione, poiché è semplicemente un'astrazione rispetto allo specifico problema di partenza.[4]

A questo punto penso sia importante definire alcuni concetti fondamentali per tale analisi.

In primo luogo, è stata data una definizione formale di equità di un algoritmo (Kleinberg, 2016), basata su diverse proprietà, in particolare:[5]

1. **Calibrazione:** la frazione di istanze rilevanti tra i risultati positivi è la stessa tra i gruppi, ossia esiste una precisione equa tra i due diversi gruppi.
2. **Bilanciamento del tasso di errore:** il tasso di errore è uguale tra i due gruppi, ossia presentano lo stesso tasso di falsi positivi e falsi negativi (volendo si può suddividere in due categorie se uno dei due errori è considerato più grave dell'altro)[6]

In secondo luogo, sono state definite diverse forme di imparità di un algoritmo (Zafar M. B., 2017):

1. **Trattamento disomogeneo:** si verifica quanto il risultato su un certo individuo si modifica modificando le informazioni sensibili dell'individuo (ed esempio la razza)
2. **Impatto disomogeneo:** si verifica quando il risultato favorisce o sfavorisce in maniera sproporzionata un certo gruppo
3. **Maltrattamenti disomogenei:** si verificano quando il tasso di classificazioni sbagliate è diverso per i due diversi gruppi

Infine, altre definizioni rilevanti sono:

1. **Tasso di riferimento:** indica la distribuzione originale della popolazione, ossia la probabilità di avere un certo risultato in ognuno dei due gruppi.

---

questo elaborato mi limiterò a considerare lo scopo principale per cui tale strumento è stato ideato, ossia la libertà vigilata e i programmi alternativi.

[3] Un'altra distinzione possibile è quella di genere.

[4] Ossia mi sembra coerente utilizzare analisi più generali di iniquità di un algoritmo rispetto due gruppi per lo specifico problema della discriminazione razziale.

[5] Ci sono anche altre definizioni importanti che però non ho approfondito in questa ricerca considerando solo quelle ritenute più significative

[6] In questo caso potremmo considerare più pesante un falso positivo rispetto ad un falso negativo perché è considerabile più grave condannare ingiustamente un innocente piuttosto che il contrario, dal mio punto di vista.

2. **Precisione generale:** indica il grado di precisione generale di un algoritmo di predizione, ossia quante predizioni corrette questo compie, indipendentemente dai gruppi di appartenenza

In base a queste definizioni possiamo considerare di avere un algoritmo equo nel caso in cui sia la calibrazione che il bilanciamento del tasso di errore siano valide *simultaneamente*, mentre nessuno dei tre effetti di disomogeneità si sia verificato.

Grazie a queste definizioni formali è più facile inquadrare il caso COMPAS: gli autori di ProPublica infatti denunciano la mancanza di bilanciamento del tasso di errore tra imputati bianchi e neri, che comporta un impatto disomogeneo su quest'ultimi[7], mentre i produttori dell'algoritmo vantano una buona calibrazione oltre che una buona precisione generale.

Il punto cruciale è che affinché uno strumento possa essere considerato equo, quindi esente da discriminazione, le due proprietà (bilanciamento e calibrazione) debbano verificarsi *simultaneamente*. Quindi per lo specifico caso di COMPAS possiamo affermare che lo strumento non sia realmente equo e sfavorisca gli imputati neri rispetto a quelli bianchi, introducendo così una forma di discriminazione razziale.

Sezione 3: EQUITA' VS PRECISIONE

Il problema principale è che questo malfunzionamento non riguarda lo specifico caso di COMPAS ma è stato dimostrato che è impossibile ottenere un algoritmo equo in presenza di gruppi con tasso di riferimento diverso (Kleinberg, 2016).

Il problema consiste proprio nel tasso di riferimento diverso: gli imputati neri hanno maggiore probabilità di recidiva rispetto ai bianchi, quindi per ottenere una buona calibrazione si ottengono tassi di errore differenti, ma nel momento in cui si diminuisce la differenza tra i tassi di errore si perde in termini di calibrazione ed anche di precisione generale.

3.1: Possibili soluzioni tecniche

Alcune soluzioni tecniche a questo problema sono state proposte, ma al momento presentano tutte forti limiti.

In particolare, una soluzione proposta da (C. Dwork, 2012) prevede l'utilizzo del dato sensibile all'interno dell'algoritmo per fissare diverse soglie di decisione per i diversi gruppi. Questo permette di ottenere dei buoni risultati, diminuendo la differenza tra i tassi di errore pur mantenendo sia una buona calibrazione che una buona precisione generale. Purtroppo, è una soluzione estrema e poco realizzabile perché appunto prevede l'utilizzo del dato sensibile (la razza in questo caso) come parametro di correzione e porterebbe ad un'altra seria di problemi (come quello dei trattamenti dei dati sensibili).

Questo modello viene superato da (Zafar M. B., 2017) grazie all'utilizzo della nozione di maltrattamenti disomogenei. Gli autori hanno definito formalmente le condizioni per cui un classificatore binario possa essere considerato esente da maltrattamenti disomogenei e hanno provato ad applicare tali condizioni anche ad un caso reale (proprio quello di COMPAS). I risultati ottenuti sono abbastanza buoni in quanto la differenza tra i tassi di errore diminuisce senza una grande perdita in calibrazione e precisione generale.

Anche questo modello, purtroppo, presenta i suoi limiti poiché non garantisce un criterio di ottimizzazione globale e utilizza approssimazioni importanti.

3.2: Conclusioni sul bilanciamento

Nonostante queste soluzioni tecniche proposte, che sicuramente andranno migliorando con gli anni, l'unica vera risposta proposta sia da (Kleinberg, 2016) che da (Berk R., 2017) è quella di effettuare un trade-off tra

[7] Come dimostrato anche da (Chouldechova, 2016),

le diverse caratteristiche dell'equità (calibrazione vs bilanciamento del tasso di errore) e in generale tra equità e precisione generale.

In particolare, (Berk R., 2017) sostiene che non sia compito di chi scrive gli algoritmi indagare questo trade-off poiché in entrambi i casi si va incontro ad un caso di iniquità e possibile discriminazione: in un caso si perde in calibrazione o precisione generale, quindi si usa uno strumento impreciso per tutti, nell'altro caso si possono avere una buona calibrazione o precisione ma un tasso di errore molto diverso, come nel caso di COMPAS, che comporta una forte discriminazione verso un gruppo della popolazione.

La mia opinione è che sia necessario porsi due domande fondamentali:[8]

- Siamo consapevoli di utilizzare uno strumento basato su questo trade-off?
- Questo trade-off è davvero l'unica causa di iniquità?

Sezione 4: SIAMO CONSAPEVOLI DI UTILIZZARE UNO STRUMENTO BASATO SU UN TRADE-OFF?

Come già detto, (Berk R., 2017) sostiene che non sia compito di chi disegna l'algoritmo preoccuparsi di questo trade-off, ma della politica e della società. Sono parzialmente d'accordo con questa affermazione ma mi sembra importante sottolineare che questi strumenti non sono in fase di creazione: molti già esistono e vengono utilizzati ampiamente.

Per quanto non stia al produttore decidere il trade-off più opportuno, essendo che la diffusione di questi strumenti è già ampia la scelta è già stata fatta e, cosa assai grave, non è stata resa nota.

Un esempio di questo problema è sicuramente il caso COMPAS, i cui produttori difendono la validità basandosi esclusivamente sulla qualità della calibrazione e della precisione generale, trascurando di menzionare il problema della differenza tra i tassi di errore.

Quindi non solo hanno valutato il trade-off da loro scelto accettabile, ma non hanno esplicitato la scelta compiuta di privilegiare la precisione e la calibrazione a scapito di un diverso tasso di errore.[9]

Penso che sia necessaria ed urgente un'operazione di informazione e consapevolizzazione riguardo a questo problema. In particolare, ritengo che un passo importante sarebbe indagare quanto l'utilizzatore di questi strumenti, ossia il giudice, sia conscio dei loro limiti. Allo stesso tempo sarebbe importante, secondo me, inserire nei criteri di valutazione e validazione degli strumenti delle informazioni riguardo al "trade-off" scelto.

Se non si può raggiungere l'equità come minimo bisogna rendere nota l'iniquità.

Sezione 5: QUESTO TRADE-OFF È DAVVERO L'UNICA CAUSA DI INIQUITÀ?

Il fatto che non si riesca a raggiungere un risultato *veramente* equo è sintomo che qualcosa non funziona.

Come è stato detto, soddisfare simultaneamente tutte le caratteristiche dell'equità, avendo popolazioni con tasso di riferimento differente è impossibile.

Uno spunto interessante riguardo a questa problematica, in particolare per le conseguenze che ne derivano, può essere evinto dallo studio di (Harcourt, 2015).

[8] Un'altra domanda importante potrebbe essere quale delle due proprietà sia più importante in ambito giudiziario ma secondo la mia opinione entrambe condurrebbero diverse forme di ingiustizia
[9] Altrimenti l'indagine di ProPublica non avrebbe suscitato tanta attenzione

Secondo questo studio, infatti, alcuni fattori considerati essenziali per predire la probabilità di recidività non sono altro che un "proxy" per la razza.

Infatti, mostra come la razza venisse spesso inclusa come fattore determinante nei primi strumenti di valutazione del rischio creati e sia stata poi successivamente esclusa dopo il cambiamento delle leggi razziali. Questo ovviamente ha portato ad un trattamento impari di fronte alla legge e ad una forte discriminazione razziale nel corso degli anni e di conseguenza ad una forte sproporzione tra la popolazione carceraria nera e quella bianca.

Per questo (Harcourt, 2015) sostiene che includere come fattore, considerato anche molto significativo, i precedenti penali e in generale la storia criminale di un imputato sia discriminatorio tanto quanto includere la razza.

Per anni è stato valutato il grado di rischio di un imputato basandosi sulla razza e questo ha portato i neri a ricevere trattamenti impari di fronte alla giustizia e quindi ad una forte sproporzione nella popolazione carceraria, ora inserendo i precedenti penali come fattore determinante non facciamo altro che aggravare questa discriminazione.

Questo studio mi ha interessata particolarmente per due motivi.

Da un lato può essere considerato un ottimo esempio del problema della scelta tra equità e precisone generale. Stando allo studio di (Harcourt, 2015) i fattori che causano iniquità sono i precedenti penali dell'imputato, quindi si potrebbe pensare di escluderli nella valutazione del rischio di recidività per ottenere un maggiore livello giustizia. Ovviamente questo comporterebbe, a mio parere, una fortissima diminuzione del grado di precisione generale dello strumento.

Dall'altro lato questo studio pone l'attenzione su quella che considero sia un'ulteriore causa di iniquità e in particolare di discriminazione: la scelta dei parametri utilizzati nel calcolo del punteggio di rischio, poiché questi parametri possono implicitamente riflettere pregiudizi e discriminazioni.[10]

Sicuramente questo problema rientra nel più ampio problema di equità presentato nella Sezione 3 e può esserne addirittura un esempio (riconsiderando l'utilizzo dei parametri "viziati" modifico la precisione della mia valutazione), ma soprattutto si focalizza su un procedimento pericoloso. Introducendo parametri intrinsecamente discriminatori trasmettiamo agli algoritmi i nostri stessi pregiudizi e discriminazioni, con la differenza che la società può acquisire consapevolezza riguardo a questi, un algoritmo li riproduce inconsciamente ed automaticamente. (e soprattutto continuerà a riprodurli finché questi parametri continueranno ad essere "viziati").

Proprio per questa caratteristica (Harcourt, 2015) suggerisce il non utilizzo in generale degli strumenti di valutazione del rischio di recidiva poiché conterranno sempre dei pregiudizi storico sociali.

La mia posizione si avvicina molto a quella di (Harcourt, 2015), sebbene creda che in futuro sarà possibile rendere questi strumenti più equi, al momento secondo me non lo sono in maniera sufficiente da rendere opportuno il loro utilizzo.

Conclusione

[10] Come visto, i precedenti penali possono contenere un forte pregiudizio razziale, ma anche altri fattori come il grado di educazione, il quartiere in cui si abita, la storia criminale familiare possono contenere questi pregiudizi.

Arianna Ricci
10536266 31/08/2020

Sono partita da un'analisi di un caso molto specifico, ossia dalla discriminazione razziale rilevata in uno strumento di predizione della recidività quale COMPAS, per riuscire a definire in maniera più generale e concettuale il problema.

Ho indagato quindi come si caratterizzi la qualità dell'equità per un algoritmo in maniera formale, riconducendo il problema della discriminazione razziale a questa più ampia cornice e rilevando che il caso di partenza non risulta equo.

A questo punto la domanda spontanea è come si possa risolvere questo problema, ma purtroppo al momento non si è trovata una soluzione tecnica valida ed applicabile che non sia un trade-off tra l'equità della valutazione e la sua precisione (che mancando condurrebbe ad altre forme di iniquità) o tra le diverse caratteristiche dell'equità.

Di nuovo sorgono spontanee altre domande ossia quanto gli utilizzatori siano consci di questo problema e se sia effettivamente l'unico problema da risolvere.

La prima risposta vuole essere più una proposta: bisognerebbe indagare il grado di consapevolezza delle caratteristiche profonde di strumenti già in uso in molti stati e soprattutto bisognerebbe rivalutare le proprietà che vengono testate e considerate per deciderne la validità.

La seconda risposta invece rileva che c'è un'implicita discriminazione contenuta in questi algoritmi, che rispecchia e deriva dalla discriminazione presente nella società (in particolare in ambito giudiziario)[11]. Questa discriminazione implicita è dovuta ai parametri che vengono utilizzati da questi strumenti, che appunto non riescono a superare i fattori di discriminazione presenti nella realtà ma anzi li assorbono e li riproducono inconsciamente nei risultati.

Si potrebbe obiettare che non sono più ingiusti della società in cui viviamo e che soprattutto sono proposti come supporto e non come sostituzione del giudizio umano. Tuttavia, ritengo che la differenza tra il pregiudizio umano e quello algoritmico risieda nel grado di consapevolezza: un essere un umano è consapevole, o può diventarlo, di essere affetto da pregiudizi, mentre un algoritmo, una volta contaminato con i nostri bias, li riprodurrà inconsapevolmente. Inoltre, questi strumenti sono presentati come un valido supporto che potrebbe addirittura diminuire la discriminazione razziale in ambito giudiziario (Skeem J., 2016), ma il loro utilizzo attuale, privo di adeguata informazione e consapevolezza, può diventare invece secondo me molto fuorviante e avallare pregiudizi già fortemente presenti.[12]

La mia conclusione è quindi quella di ripensare e limitare l'utilizzo di questi strumenti e la loro validità.

Poiché non possiamo tecnicamente garantirne l'assoluta equità e poiché appunto consolidano e trasmettono inconsapevolmente una discriminazione della società, credo non possano rientrare in quella cornice di fiducia proposta da (Scantamburlo T., 2018) e quindi non siano, per ora, di reale supporto alla decisione umana.[13]

Tuttavia, poiché l'utilizzo di questi strumenti si sta già diffondendo, credo che almeno debba essere accompagnato da una forte presa di coscienza e di consapevolezza. Sono apprezzabili gli sforzi compiuti per trovare una soluzione al problema del trade-off, ma finché queste soluzioni non saranno realmente efficaci,

[11] Prima fra tutte la disparità nella popolazione carceraria

[12] Un giudice che tenta di superare un possibile pregiudizio razziale, vedendo i punteggi dei neri sempre più alti (erroneamente) di quelli da lui previsti e avendo fiducia nell'imparzialità dello strumento, potrebbe essere indotto a riprendere un pregiudizio che aveva superato.

[13] Infatti, rischiano di aggravare la discriminazione già presente aggiungendo un ulteriore elemento discriminante, per di più prodotto da un algoritmo che non ha consapevolezza di questo bias, diversamente dagli esseri umani.

gli utilizzatori dovrebbero essere informati dei limiti intrinsechi a questi strumenti e delle scelte di design compiute.

Riferimenti

Angwin, J. L. (2016). How We Analyzed the COMPAS Recidivism Algorithm.

Angwin, J. L. (2016). Machine Bias.

Berk R., H. H. (2017). Fairness in Criminal Justice Risk Assessments:.

C. Dwork, M. H. (2012). Fairness Through Awareness.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

Flores A. W., L. C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.".

Harcourt, B. (2015). Risk as a proxy for race: The dangers of risk assessment.

Kleinberg, J. M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores.

Scantamburlo T., C. A. (2018). MACHINE DECISIONS AND HUMAN CONSEQUENCES.

Skeem J., L. C. (2016). Risk, Race, & Recidivism: Predictive Bias and Disparate Impact.

Zafar M. B., V. I. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment.