Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria dell'Informazione

Corso di laurea in Ingegneria Informatica

Misurare la fairness negli algoritmi: il caso di studio COMPAS

Candidata:
Giulia Bonato

Relatore:
Prof. Antonio Rodà

Correlatrice:
Prof.ssa Silvana Badaloni

Anno Accademico 2023/2024

Data di laurea: 15 luglio 2024

Abstract

La pervasività degli algoritmi nelle decisioni quotidiane ha suscitato crescenti preoccupazioni, si teme infatti che alcuni di essi siano affetti da pregiudizi che possano promuovere discriminazioni e ingiustizie.
Questa tesi si propone di replicare ed ampliare le analisi condotte in un noto studio di ProPublica del 2016 sull'algoritmo COMPAS, ampiamente utilizzato nel sistema giudiziario statunitense per predire la recidiva criminale. Inoltre, verrà utilizzato il toolkit di audit Aequitas per esplorare e confrontare le implicazioni di giustizia algoritmica nei risultati prodotti da COMPAS.
Questo lavoro non offre solo un'analisi critica della fairness di COMPAS, ma dimostra anche l'efficacia di strumenti come Aequitas nel promuovere una maggiore trasparenza e responsabilità nello sviluppare algoritmi decisionali e nella loro diffusione.
Infine, questo studio vuole proporsi come un contributo per perseguire una società in cui l'equità e la giustizia guidino l'evoluzione della tecnologia.

Indice

Capitolo 1

Introduzione

Questo elaborato è suddiviso in 7 capitoli principali, ciascuno dei quali esplora un aspetto specifico del tema in esame. Nel Capitolo 1, sono riportate le definizioni del concetto di Fairness negli algoritmi e di alcune metriche significative per misurarla. Nel Capitolo 2, ho inserito una trattazione in cui elenco i principali tipi di bias di cui possono essere affetti i dati. All'interno del Capitolo 3, dapprima viene introdotto il sistema di valutazione COMPAS ed il suo funzionamento, in seguito le analisi prodotte da ProPublica alle quali segue la risposta di Northpointe di cui ho riportato i punti salienti. Nel Capitolo 4, ho riprodotto le analisi svolte da ProPublica e le ho ampliate producendo un'analisi intersezionale. Nel Capitolo 5, illustro le funzionalità di Aequitas, strumento per la misurazione della fairness e per la valutazione della presenza di bias all'interno degli algoritmi. All'interno del Capitolo 6, ho riportato e commentato i risultati raccolti nel "The Bias Report" prodotto da Aequitas.
Infine nel Capitolo 7, ho confrontato i risultati ottenuti da ProPublica ed Aequitas verificando se fossero coerenti tra loro.

1.1 Fairness negli algoritmi

La lotta contro i pregiudizi e le discriminazioni è un tema che affonda le sue radici in campi quali la filosofia e la psicologia fino ad arrivare ad applicazioni più recenti come nell'ambito del machine learning. Tuttavia, per poter combattere tale fenomeno e raggiungere uno stato generalizzato di equità (meglio nota con il termine di fairness in letteratura), è necessario prima definire il concetto di fairness.
La filosofia e la psicologia hanno cercato di definire il concetto di fairness molto prima dell'informatica. Il fatto che non esista una definizione univoca di fairness dimostra la difficoltà nel risolvere questo problema. Infatti all'interno della società coesistono visioni e prospettive diverse rispetto tale tema che rendono ancora più difficile individuare una definizione universale riconosciuta da tutti. In generale, la fairness è l'assenza di pregiudizi o favoritismi nei confronti di un individuo o di un gruppo in base alle loro caratteristiche intrinseche o acquisite nel contesto del processo decisionale.

Tenendo conto di queste sfide, sono state proposte molte definizioni di fairness per affrontare i diversi problemi che sorgono a causa della presenza di algoritmi affetti da pregiudizi (meglio noti sotto il nome di bias in letteratura).

1.2 Metriche per misurare la Fairness

In questo paragrafo verranno presentate alcune delle definizioni attribuite al termine fairness. Pertanto, prima di iniziare con la trattazione è necessario fornire alcune definizioni di riferimento: *Gruppo protetto* - gli individui accomunati da alcuni attributi protetti quali genere, etnia, estrazione sociale, religione, ecc., che non possono essere adottati per prendere decisioni, in quanto potrebbero essere fonte di discriminazione. *Classe positiva* - si riferisce alla classe che il modello è interessato a rilevare o prevedere. Rappresenta la presenza o il verificarsi della caratteristica o della condizione specifica che il modello sta cercando di identificare.

1. *Equalized Odds* - La definizione di probabilità equalizzata [1] afferma che "un predittore $\hat{Y}$ soddisfa probabilità equalizzate rispetto all'attributo protetto A e al risultato Y, se $\hat{Y}$ e A sono indipendenti condizionatamente da Y.
   $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$". Ciò significa che la probabilità che a una persona della classe positiva venga assegnato correttamente un esito positivo e la probabilità che a una persona della classe negativa venga assegnato erroneamente un esito positivo dovrebbero essere entrambe uguali per i membri del gruppo protetto e non protetto [2]. In altre parole, la definizione di probabilità equalizzate afferma che i gruppi protetti e non protetti devono avere tassi uguali di veri positivi e falsi positivi.

2. *Equal Opportunity* - "Un predittore binario $\hat{Y}$ soddisfa la pari opportunità rispetto ad A e Y se $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$" [1]. Ciò significa che la probabilità che una persona appartenente a una classe positiva sia assegnata a un risultato positivo deve essere uguale sia per i membri del gruppo protetto che per quelli non protetti (donne e uomini) [2]. In pratica, la definizione di pari opportunità afferma che i gruppi protetti e non protetti dovrebbero avere uguali tassi di veri positivi.

3. *Demographic Parity* - Conosciuta anche come parità statistica o *Statistical Parity*. "Un predittore $\hat{Y}$ soddisfa la parità demografica se $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$" [3, 4]. La probabilità di un esito positivo [2] dovrebbe essere la stessa indipendentemente dal fatto che la persona faccia parte del gruppo protetto.

4. *Fairness through Awareness* - "Un algoritmo è equo se fornisce previsioni simili a individui simili" [3, 4]. In altre parole, due individui che sono simili rispetto a una metrica di somiglianza (distanza inversa) definita per un particolare compito dovrebbero ricevere un risultato simile.

5. *Fairness through Unawareness* - "Un algoritmo è equo fintanto che gli attributi protetti A non vengono utilizzati esplicitamente nel processo decisionale" [5, 4].

6. *Treatment Equality* - "L'uguaglianza di trattamento si ottiene quando il rapporto tra falsi negativi e falsi positivi è lo stesso per entrambe le categorie di gruppi protetti" [6].

7. *Test Fairness* - "Un punteggio $S = S(x)$ è test fair (ben calibrato) se riflette la stessa probabilità di recidiva indipendentemente dall'appartenenza al gruppo dell'individuo, R. Ovvero, se per tutti i valori di s, $P(Y = 1|S = s, R = b) = P(Y = 1|S = s, R = w)$" [7]. In altre parole, la definizione di equità del test afferma che per qualsiasi punteggio di probabilità previsto S, le persone appartenenti ai gruppi protetti e non protetti devono avere la stessa probabilità di appartenere correttamente alla classe positiva [2].

8. *Counterfactual Fairness* - "Il predittore $\hat{Y}$ è controfattualmente equo se, in qualsiasi contesto $X = x$ e $A = a$, $P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)$, per tutti gli y e per qualsiasi valore a' attribuibile da A" [4]. La definizione di equità controfattuale si basa sull'"intuizione che una decisione è equa nei confronti di un individuo se è la stessa sia nel mondo reale sia in un mondo controfattuale in cui l'individuo appartiene a un gruppo demografico diverso".

9. *Fairness in Relational Domains* - "Una nozione di equità in grado di catturare la struttura relazionale di un dominio, non solo prendendo in considerazione gli attributi degli individui, ma anche le connessioni sociali, organizzative e di altro tipo tra gli individui" [8].

10. *Conditional Statistical Parity* - Per un insieme di fattori legittimi L, il predittore $\hat{Y}$ soddisfa la parità statistica condizionata se $P(\hat{Y}|L = 1, A = 0) = P(\hat{Y}|L = 1, A = 1)$ [9]. La parità statistica condizionata stabilisce che le persone dei gruppi protetti e non protetti (donne e uomini) devono avere la stessa probabilità di essere assegnate a un esito positivo dato un insieme di fattori legittimi L [2].

Le definizioni citate possono essere classificate in tre categorie:

- *Individual Fairness*: fornire previsioni simili a individui simili.
- *Group Fairness*: trattare gruppi diversi nello/allo stesso modo.
- *Subgroup Fairness*: l'equità di sottogruppo mira a ottenere le migliori proprietà delle nozioni di equità di gruppo e individuale.

Capitolo 2

Algoritmi affetti da bias

2.1 Tipi di bias

La maggior parte dei sistemi e degli algoritmi di intelligenza artificiale sono basati sui dati e necessitano di dati per essere addestrati. I dati sono quindi strettamente legati alla funzionalità di questi algoritmi e sistemi. Nel caso in cui i dati di addestramento sottostanti contengano distorsioni, gli algoritmi addestrati su di essi apprenderanno tali distorsioni e li rifletteranno nelle loro previsioni. Di conseguenza, i pregiudizi esistenti nei dati possono influenzare gli algoritmi che li utilizzano, producendo risultati distorti. Gli algoritmi possono persino amplificare e perpetuare i pregiudizi esistenti nei dati. Inoltre, gli algoritmi stessi possono mostrare un comportamento distorto a causa di alcune scelte progettuali, anche se i dati in sé non sono distorti. I risultati di questi algoritmi distorti possono poi essere immessi nei sistemi del mondo reale e influenzare le decisioni degli utenti, il che si tradurrà in dati più distorti per l'addestramento di algoritmi futuri. Il ciclo che cattura questo feedback tra i pregiudizi nei dati, gli algoritmi e l'interazione con l'utente è mostrato nella figura che segue [10].

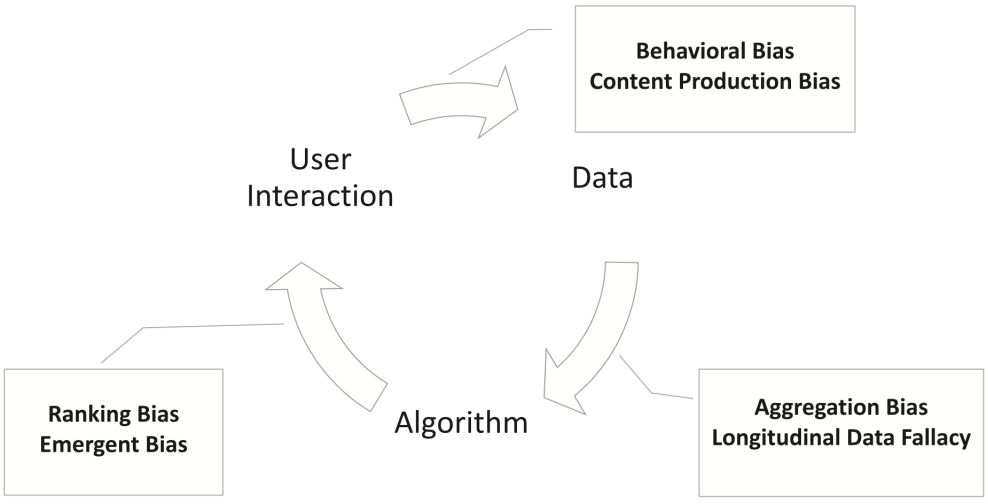

Figura 2.1: Ciclo di feedback tra i pregiudizi presenti nei dati, gli algoritmi e l'interazione con l'utente [10]

I bias possono esistere in molte forme, alcune delle quali possono portare all'iniquità in diversi ambiti in futuro se non vengono individuati tempestivamente.

2.1.1 Dai dati all'algoritmo

In questa sezione si parla di distorsioni nei dati che, se utilizzati dagli algoritmi di addestramento in ML, potrebbero portare a risultati algoritmici distorti [10].

1. **Bias di misurazione** - I bias di misura derivano dal modo in cui scegliamo, utilizziamo e misuriamo particolari caratteristiche [11]. Un esempio di questo tipo di bias è stato osservato nello strumento di previsione del rischio di recidiva COMPAS, che sarà oggetto di approfondimento nei/dei prossimi capitoli, in cui gli arresti precedenti e gli arresti di amici/familiari sono stati utilizzati come variabili proxy per misurare il livello di "rischiosità" o "criminalità", che di per sé possono essere considerati proxy misurati in modo errato. Ciò è in parte dovuto al fatto che le comunità di minoranza sono controllate e sorvegliate più frequentemente, quindi hanno tassi di arresto più elevati. Tuttavia, non si deve concludere che, poiché le persone provenienti da gruppi di minoranza hanno tassi di arresto più elevati, siano quindi più pericolose, poiché esiste una differenza nel modo in cui questi gruppi vengono valutati e controllati.

2. **Bias delle variabili omesse** - L'omissione di variabili si verifica quando una o più variabili importanti vengono escluse dal modello.

3. **Bias di rappresentazione** - Il bias di rappresentazione deriva dal modo in cui campioniamo una popolazione durante il processo di raccolta dei dati [11]. I campioni non rappresentativi mancano della diversità della popolazione, con sottogruppi mancanti e altre anomalie. La mancanza di diversità geografica in dataset come ImageNet, ad esempio, si traduce in un pregiudizio dimostrabile verso le culture occidentali.

4. **Bias di aggregazione** - Il bias di aggregazione si verifica quando si traggono conclusioni errate sugli individui osservando l'intera popolazione. Un modello che ignora le differenze individuali probabilmente non sarà adatto a tutti i gruppi etnici e di genere della popolazione [11]. Questo è vero anche quando sono rappresentati in modo uguale nei dati di addestramento. Qualsiasi ipotesi generale sui sottogruppi all'interno della popolazione può dare luogo a pregiudizi di aggregazione.

    (a) **Paradosso di Simpson** - Il paradosso di Simpson è un tipo di bias di aggregazione che si verifica nell'analisi di dati eterogenei [12]. Il paradosso si verifica quando un'associazione osservata in dati aggregati scompare o si inverte quando gli stessi dati vengono disaggregati nei sottogruppi sottostanti.

   (b) **Unità areale modificabile** - Il problema dell'unità areale modificabile è un errore statistico nell'analisi geospaziale che si verifica quando si modellano i dati a diversi livelli di aggregazione spaziale. Questo bias si traduce in tendenze diverse apprese quando i dati sono aggregati a scale spaziali diverse.

5. **Bias di campionamento** - Il bias di campionamento è simile al bias di rappresentazione e consiste in un campionamento non casuale di sottogruppi. Come conseguenza del sampling bias, le tendenze stimate per una popolazione possono non essere generalizzate ai dati raccolti da una nuova popolazione.

6. **Fallacia dei dati longitudinali** - I ricercatori che analizzano i dati temporali devono utilizzare l'analisi longitudinale per seguire le coorti nel tempo e imparare il loro comportamento. Invece, i dati temporali sono spesso modellati utilizzando l'analisi trasversale, che combina coorti diverse in un unico punto temporale. L'eterogeneità delle coorti può influenzare l'analisi trasversale, portando a conclusioni diverse rispetto all'analisi longitudinale.

7. **Linking bias** - Il linking bias si verifica quando gli attributi della rete ottenuti dalle connessioni, dalle attività o dalle interazioni degli utenti differiscono e travisano il vero comportamento degli utenti.

2.1.2 Dall'algoritmo all'utente

In questa sezione si parla di pregiudizi che sono il risultato di risultati algoritmici e che di conseguenza influenzano il comportamento dell'utente.

1. **Bias algoritmico** - Si parla di bias algoritmico quando il bias non è presente nei dati in input e viene aggiunto esclusivamente dall'algoritmo [13]. Le scelte di progettazione dell'algoritmo, come l'uso di determinate funzioni di ottimizzazione, le regolarizzazioni, le scelte di applicare i modelli di regressione ai dati nel loro complesso o di considerare sottogruppi e l'uso generale di stimatori statisticamente distorti negli algoritmi, possono contribuire a decisioni algoritmiche distorte che possono influenzare il risultato degli algoritmi.

2. **Pregiudizi dell'interazione con l'utente** - Il pregiudizio dell'interazione con l'utente è un tipo di pregiudizio che può essere osservato non solo sul Web, ma anche innescato da due fonti: l'interfaccia dell'utente e l'utente stesso, che impone il suo comportamento e la sua interazione auto-selezionata e distorta [13]. Questo tipo di pregiudizio può essere influenzato da altri tipi e sottotipi, come i pregiudizi di presentazione e di classificazione.

   (a) **Bias di presentazione** - Il bias di presentazione è il risultato del modo in cui le informazioni vengono presentate [13]. Ad esempio, sul Web gli

utenti possono cliccare solo sui contenuti che vedono, quindi i contenuti visti ricevono click, mentre tutto il resto non riceve click. Potrebbe anche accadere che l'utente non veda tutte le informazioni presenti sul Web [13].

(b) **Bias di classificazione** - L'idea che i risultati in cima alla classifica siano i più rilevanti e importanti, attirerà un maggior numero di click rispetto agli altri. Questo pregiudizio riguarda principalmente i motori di ricerca.

3. **Bias di popolarità** - Gli articoli più popolari tendono a essere esposti di più. Tuttavia, le metriche di popolarità sono soggette a manipolazione, ad esempio da parte di recensioni false o bot sociali. Ad esempio, questo tipo di bias può essere riscontrato nei motori di ricerca o nei sistemi di raccomandazione, dove gli oggetti più popolari vengono presentati maggiormente al pubblico.

4. **Bias emergenti** - I bias emergenti si verificano in seguito all'uso e all'interazione con gli utenti reali. Questi bias si verificano in seguito a cambiamenti nella popolazione, nei valori culturali o nelle conoscenze della società, di solito qualche tempo dopo il completamento della progettazione.

5. **Bias di valutazione** - I bias di valutazione si verificano durante la valutazione dei modelli [11]. Questo include l'uso di benchmark inappropriati e sproporzionati per la valutazione delle applicazioni.

2.1.3 Dall'utente ai dati

Molte fonti di dati utilizzate per l'addestramento di modelli ML sono generate dagli utenti. Qualsiasi pregiudizio intrinseco degli utenti potrebbe riflettersi nei dati da loro generati. Di seguito elenco alcuni tipi importanti di tali pregiudizi.

1. **Bias storico** - I bias storici sono i pregiudizi e i problemi socio-tecnici già esistenti nel mondo che possono insinuarsi nel processo di generazione dei dati anche con un campionamento e una selezione delle caratteristiche perfetti [11].

2. **Bias della popolazione** - I bias di popolazione si verificano quando le statistiche, i dati demografici, i rappresentanti e le caratteristiche degli utenti sono diversi nella popolazione di utenti della piattaforma rispetto alla popolazione target originale. I bias di popolazione creano dati non rappresentativi.

3. **Bias di autoselezione** - Il bias di autoselezione è un sottotipo di bias di selezione o campionamento in cui i soggetti della ricerca selezionano se stessi. Un esempio di questo tipo di bias può essere osservato in un sondaggio d'opinione per misurare l'entusiasmo per un candidato politico, dove i sostenitori più entusiasti hanno maggiori probabilità di completare il sondaggio.

4. **Bias sociale** - I bias sociali si verificano quando le azioni degli altri influenzano il nostro giudizio [13].

5. **Bias comportamentale** - I bias comportamentali derivano dal diverso comportamento degli utenti tra piattaforme, contesti o insiemi di dati diversi.

6. **Bias temporale** - I bias temporali derivano dalle differenze di popolazioni e comportamenti nel corso del tempo.

7. **Bias di produzione dei contenuti** - I bias di produzione dei contenuti derivano da differenze strutturali, lessicali, semantiche e sintattiche nei contenuti generati dagli utenti.

I lavori esistenti cercano di classificare queste definizioni di bias in gruppi, come le definizioni che rientrano esclusivamente nei dati o nell'interazione con l'utente.
Tuttavia, a causa dell'esistenza del fenomeno del feedback loop (vedi Figura 2.1), queste definizioni sono interconnesse e bisogna considerare come si influenzano a vicenda in questo ciclo e affrontarle di conseguenza.

Capitolo 3

COMPAS Core

3.1 COMPAS - Risk Assessment Tool

Nei sistemi di giustizia penale sovraccarichi e affollati, la rapidità, l'efficienza, la facilità di gestione e la chiara organizzazione dei dati chiave sui rischi/bisogni sono fondamentali. L'algoritmo COMPAS è stato presentato per la prima volta dalla compagnia Northpointe nel marzo 2015 con lo scopo di fornire un valido strumento per ottimizzare questi fattori. COMPAS affronta questo compromesso in diversi modi: fornisce un insieme completo di fattori di rischio chiave emersi dalla recente letteratura criminologica e consente la personalizzazione del software. Pertanto, la facilità d'uso, la gestione efficiente ed efficace del tempo e le considerazioni sulla gestione dei casi che sono critiche per le migliori pratiche nel campo della giustizia penale possono essere raggiunti attraverso COMPAS, afferma la società [14].

3.2 Punteggi COMPAS

Il sistema di valutazione COMPAS è composto da scale di rischio predittive per la previsione del rischio e da scale di bisogno separate per l'identificazione dei bisogni del programma nei settori dell'occupazione, dell'alloggio e dell'abuso di sostanze e altri.
Le agenzie si attengono comunemente al principio del rischio per indirizzare ai programmi di trattamento individui che presentano un elevato punteggio di rischio di recidiva e un elevato bisogno di trattamento (ad esempio, un elevato consumo di sostanze).

3.2.1 AIPIE

L'interpretazione e gli eventi correlati alla gestione dei casi possono essere una serie di attività complesse per i professionisti. Un modello che aiuta a spiegare le procedure della pratica basata sull'evidenza è noto come AIPIE. Il modello AIPIE è strutturato in modo tale che l'informazione innesca decisioni che innescano azioni.

A = Valutazione (attraverso COMPAS o un altro strumento)
I = Interpretazione dei risultati
P = Pianificare, creare un piano d'azione basato sulle informazioni raccolte
I = Implementare o attuare il piano
E = Valutare i risultati delle azioni e degli esiti

3.2.2 Conversione dei punteggi dalla scala grezza in punteggi decili

I punteggi della scala COMPAS vengono trasformati in punteggi decili. I decili si ottengono classificando i punteggi della scala di un gruppo normativo in ordine ascendente e poi dividendo questi punteggi in dieci gruppi di uguali dimensioni. I decili vanno da 1 (minimo) a 10 (massimo) [14].

In generale, il grado di decile ha la seguente interpretazione:

- $1-4$: il punteggio della scala è basso rispetto agli altri autori di reato del gruppo di riferimento.
- $5-7$: il punteggio della scala è medio rispetto agli altri autori di reato del gruppo di riferimento.
- $8-10$: il punteggio della scala è alto rispetto agli altri autori di reato del gruppo di riferimento.

Type 1	Low (1-4)	Medium (5-7)	High (8-10)
Type 2	Unlikely (1-2)	Probable (3-4)	Highly Probable (5-10)
Type 3	Unlikely (1-5)	Probable (6-7)	Highly Probable (8-10)
Type 4	Unlikely (1-4)	Probable (5-7)	Highly Probable (8-10)

Figura 3.1: Fasce dei punteggi decili COMPAS/Punti di taglio [14]

È importante notare che i punteggi decili possono essere interpretati solo in senso relativo e sono sempre legati al gruppo di riferimento. È inoltre significativo osservare che per alcune scale non è sempre possibile suddividere il campione in dieci gruppi di dimensioni esattamente uguali. Per questo motivo, per alcune scale è stato necessario saltare alcuni punteggi decili. Quando non è stato possibile suddividere il campione in dieci gruppi, è stato utilizzato un algoritmo per punti di taglio che dividessero gli autori di reato nel maggior numero possibile di gruppi di dimensioni approssimativamente uguali e che utilizzassero l'intera gamma di valori decili (cioè, da 1 a 10).

3.2.3 Validità predittiva delle scale di rischio COMPAS

Northpointe, la società che ha sviluppato COMPAS Core, ha documentato i risultati delle ricerche condotte da più studi per dimostrare che il loro prodotto è affidabile, che le sue scale di misurazione dei bisogni hanno validità di costrutto e si comportano in modo coerente e che le sue scale di rischio hanno validità predittiva. Questo è ciò che sostiene l'azienda alla luce di alcuni studi che dimostrano la validità del loro algoritmo. Se si desidera consultare tali studi si rimanda alla sezione dedicata presente in [14].

3.3 Analisi ProPublica di COMPAS

3.3.1 Background

In un pomeriggio di primavera del 2014, Brisha Borden era in ritardo per andare a prendere la sua sorellina a scuola quando notò una bicicletta da bambino senza lucchetto e un monopattino lungo la strada. Borden e un'amica hanno afferrato la bicicletta e il monopattino e hanno cercato di guidarli lungo la strada nel sobborgo di Fort Lauderdale, Coral Springs. Proprio mentre le diciottenni si rendevano conto di essere troppo grandi per i piccoli mezzi di trasporto, che appartenevano a un bambino di 6 anni, una donna è accorsa dicendo: "Sono di mio figlio!". Borden e la sua amica hanno immediatamente lasciato cadere la bicicletta e il monopattino e si sono allontanate.
Ma era troppo tardi: un vicino che aveva assistito al furto aveva già chiamato la polizia. Borden e la sua amica sono state arrestate e accusate di furto con scasso e furto di minore entità per gli oggetti, che avevano un valore complessivo di 80 dollari [15].

Confrontate il loro crimine con uno simile: l'estate precedente, Vernon Prater, 41 anni, era stato arrestato per aver rubato degli utensili per un valore complessivo di 86,35 dollari da un vicino negozio di articoli per la casa. Prater era già stato condannato per rapina a mano armata e tentata rapina a mano armata, per cui aveva scontato cinque anni di carcere, oltre a un'altra accusa di rapina a mano armata. Anche Borden aveva dei precedenti, ma per reati minori commessi quando era minorenne [15].

Eppure è successo qualcosa di strano quando Borden e Prater sono stati registrati in carcere: l'algoritmo COMPAS ha fornito un punteggio che prevedeva la probabilità che ciascuno di loro commettesse un crimine in futuro. Borden, che è nera, è stata giudicata ad alto rischio. Prater, che è bianco, è stato giudicato a basso rischio.
Due anni dopo, è noto che l'algoritmo aveva sbagliato le sue valutazioni. Borden non è stata accusata di nuovi reati. Prater sta scontando una pena detentiva di otto anni per essersi introdotto in un magazzino e aver rubato migliaia di dollari di materiale elettronico [15].
Punteggi come questo, noti come valutazioni del rischio, sono sempre più comuni nei

tribunali americani. Vengono utilizzati per prendere decisioni su chi può essere rimesso in libertà in ogni fase del sistema giudiziario penale, dall'assegnazione di cauzioni, come nel caso di Fort Lauderdale, a decisioni ancora più fondamentali sulla libertà degli imputati. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington e Wisconsin, i risultati di tali valutazioni vengono forniti ai giudici durante la sentenza penale [15].

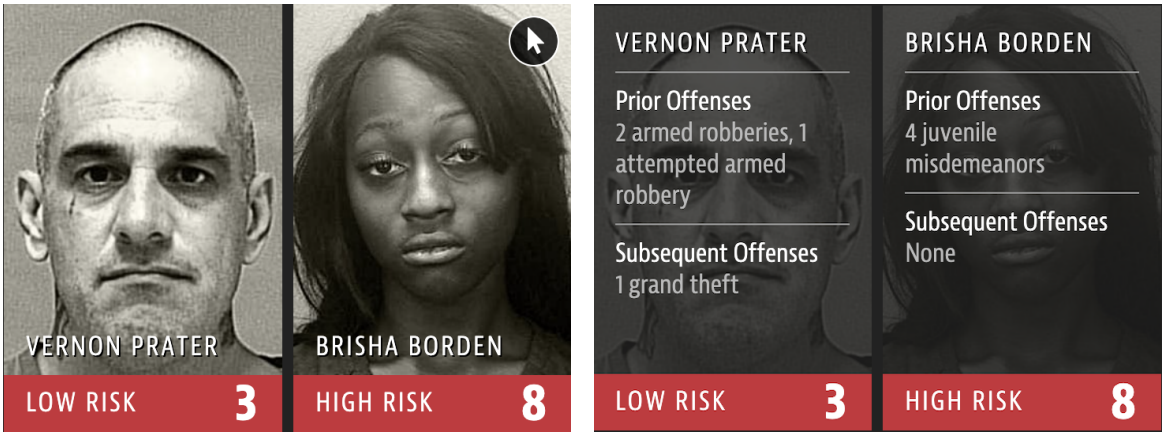

Figura 3.2: Borden è stata classificata ad alto rischio di reati futuri dopo che, insieme a un'amica, ha preso la bicicletta e il monopattino di un bambino che si trovavano all'aperto. Non ha commesso un nuovo reato.

Nel 2014, l'allora procuratore generale degli Stati Uniti Eric Holder aveva avvertito che i punteggi di rischio potevano introdurre pregiudizi nei tribunali. Ha chiesto alla Commissione per le sentenze degli Stati Uniti di studiarne l'uso. "Sebbene queste misure siano state elaborate con le migliori intenzioni, temo che inavvertitamente minino i nostri sforzi per garantire una giustizia equa e personalizzata", ha dichiarato, aggiungendo che "potrebbero esacerbare disparità ingiustificate e ingiuste che sono già troppo comuni nel nostro sistema di giustizia penale e nella nostra società".
La Commissione per le sentenze non ha tuttavia avviato uno studio sui punteggi di rischio. Lo ha fatto ProPublica, nell'ambito di un esame più ampio del potente effetto, in gran parte nascosto, degli algoritmi nella vita americana [15].

3.3.2 Analisi condotte

ProPublica si è proposta di valutare l'accuratezza di fondo dell'algoritmo di recidiva COMPAS (acronimo di Correctional Offender Management Profiling for Alternative Sanctions) di Northpointe Inc. e di verificare se l'algoritmo fosse prevenuto nei confronti di alcuni gruppi. Le loro analisi hanno rilevato che gli imputati neri avevano molte più probabilità di quelli bianchi di essere erroneamente giudicati a rischio di recidiva, mentre gli imputati bianchi avevano più probabilità di quelli neri di essere erroneamente segnalati come a basso rischio. Durante le analisi sono stati esaminati più di 10.000 imputati della contea di Broward, in Florida, e sono stati confrontati i tassi di

recidiva previsti con quelli effettivamente verificatisi in un periodo di due anni. Quando la maggior parte degli imputati viene registrata in carcere, risponde a un questionario COMPAS. Le loro risposte vengono inserite nel software COMPAS per generare diversi punteggi, tra cui le previsioni del "rischio di recidiva" e del "rischio di recidiva violenta".

Hanno confrontato le categorie di rischio di recidiva previste dallo strumento COMPAS con i tassi effettivi di recidiva degli imputati nei due anni successivi al punteggio ed è stato scoperto che il punteggio prevedeva correttamente la recidiva di un condannato nel 61% dei casi, ma era corretto solo nelle previsioni di recidiva violenta nel 20% dei casi. Per quanto riguarda la previsione della recidiva, l'algoritmo ha previsto correttamente la recidiva per gli imputati bianchi e neri più o meno alla stessa percentuale (59% per gli imputati bianchi e 63% per gli imputati neri), ma ha commesso errori in modi molto diversi. L'algoritmo sbaglia a classificare in modo diverso gli imputati bianchi e neri quando viene esaminato in un periodo di follow-up di due anni.

3.3.3 Acquisizione dei dati

Attraverso una richiesta di documenti pubblici, ProPublica ha ottenuto due anni di punteggi COMPAS dall'ufficio dello sceriffo della contea di Broward, in Florida. Inoltre ha ricevuto i dati di tutte le 18.610 persone a cui è stato assegnato il punteggio nel 2013 e nel 2014.

Poiché la Contea di Broward utilizza il punteggio principalmente per determinare se rilasciare o detenere un imputato prima del processo, sono stati scartati i punteggi che erano stati valutati in fase di libertà vigilata o altre fasi del sistema di giustizia penale. Sono rimaste 11.757 persone valutate in fase preprocessuale.
Ogni imputato in attesa di giudizio ha ricevuto almeno tre punteggi COMPAS: "Rischio di recidiva", "Rischio di violenza" e "Rischio di mancata comparizione". I punteggi COMPAS per ogni imputato variavano da 1 a 10, con dieci che rappresentava il rischio più elevato. I punteggi da 1 a 4 sono stati etichettati da COMPAS come "basso", da 5 a 7 come "medio" e da 8 a 10 come "alto".

Partendo dal database dei punteggi COMPAS, hanno costruito un profilo della storia criminale di ogni persona, sia prima che dopo il punteggio. Dopo aver raccolto i precedenti penali pubblici dal sito web dell'ufficio della contea di Broward fino all'1 aprile 2016, hanno osservato che in media gli imputati del loro set di dati non sono stati incarcerati per 622,87 giorni (sd: 329,19). In seguito, hanno confrontato i registri penali con i registri COMPAS utilizzando il nome, il cognome e la data di nascita di una persona. Si tratta della stessa tecnica utilizzata nello studio di validazione di COMPAS della contea di Broward condotto dai ricercatori della Florida State University nel 2010. Hanno scaricato circa 80.000 registri penali dal sito web del Broward

County Clerk's Office. Per determinare l'etnia, sono state adottate le classificazioni delle etnie utilizzate dall'Ufficio dello sceriffo della contea di Broward, che identifica gli imputati come neri, bianchi, ispanici, asiatici e nativi americani. In 343 casi, l'etnia è stata indicata come Altro.

Successivamente hanno compilato il record di incarcerazione di ogni persona, dopo aver ricevuto i registri delle carceri dall'ufficio dello sceriffo della contea di Broward dal gennaio 2013 all'aprile 2016 e altri registri pubblici delle carcerazioni disponibili sul sito web del Dipartimento di correzione della Florida. Tuttavia, è stato riscontrato che a volte i nomi o le date di nascita delle persone sono stati inseriti in modo errato in alcuni registri, il che ha portato a corrispondenze errate tra il punteggio COMPAS di un individuo e i suoi precedenti penali. Di conseguenza, si è cercato di determinare il numero di record interessati: su un campione casuale di 400 casi, è stato riscontrato un tasso di errore del 3.75%.

3.3.4 Definizione di recidiva

La definizione di recidiva è stata fondamentale per le analisi condotte da ProPublica. In particolare, hanno interpretato questo concetto come un reato che ha portato alla registrazione in carcere e che è avvenuto dopo il reato per il quale la persona ha ottenuto il punteggio COMPAS. Tuttavia, non era sempre chiaro quale caso penale fosse associato al punteggio COMPAS di un individuo. Per abbinare i punteggi COMPAS ai casi che li accompagnano, sono stati presi in considerazione i casi con date di arresto o di imputazione entro 30 giorni dall'esecuzione della valutazione COMPAS. In alcuni casi, non è stato possibile trovare accuse corrispondenti ai punteggi COMPAS per cui sono stati esclusi dall'analisi.
Si è cercato, inoltre, di determinare se una persona fosse stata accusata di un nuovo reato successivo a quello per cui era stata sottoposta a screening COMPAS. Non sono state considerate le multe e alcune violazioni di ordinanze comunali di entità minore come recidiva. Allo stesso modo non sono stati considerati recidivi coloro che sono stati arrestati per non essersi presentati alle udienze in tribunale, né coloro che sono stati successivamente accusati di un reato avvenuto prima dello screening COMPAS.
Per la recidiva violenta, è stata adottata la definizione di crimine violento dell'FBI, una categoria che comprende omicidio, omicidio colposo, stupro, rapina e aggressione aggravata. Mentre, per il resto dell'analisi la recidiva è stata definita come un nuovo arresto entro due anni come confermato d'altra parte da un recente studio della Commissione per le sentenze degli Stati Uniti.

3.3.5 Risultati ottenuti

Per riassumere le analisi di ProPublica hanno rilevato che:

- Gli imputati neri sono stati spesso previsti a rischio di recidiva più elevato di quanto non fossero in realtà. La nostra analisi ha rilevato che gli imputati neri che non hanno commesso recidiva nell'arco di due anni avevano una probabilità quasi doppia di essere classificati erroneamente come ad alto rischio rispetto alle loro controparti bianche (45% contro 23%).

- Gli imputati bianchi erano spesso previsti come meno rischiosi di quanto non fossero. La nostra analisi ha rilevato che gli imputati bianchi che hanno recidivato nei due anni successivi sono stati erroneamente etichettati come a basso rischio quasi due volte più spesso dei recidivi neri (48% contro 28%).

- L'analisi ha anche mostrato che, anche controllando i reati precedenti, la recidiva futura, l'età e il sesso, gli imputati neri avevano il 45% in più di probabilità di essere assegnati a punteggi di rischio più elevati rispetto agli imputati bianchi.

- Gli imputati neri avevano anche una probabilità doppia rispetto a quelli bianchi di essere classificati erroneamente come a maggior rischio di recidiva violenta. Inoltre, gli imputati bianchi recidivi violenti avevano il 63% in più di probabilità di essere classificati erroneamente a basso rischio di recidiva violenta, rispetto agli imputati neri recidivi violenti.

- L'analisi della recidiva violenta ha mostrato che anche controllando i reati precedenti, la recidiva futura, l'età e il sesso, gli imputati neri avevano il 77% in più di probabilità di essere assegnati a punteggi di rischio più alti rispetto agli imputati bianchi.

3.4 Risposta di Northpointe a ProPublica

3.4.1 Dimostrare l'equità dell'accuratezza e la parità predittiva

In questa sezione, ho riportato la risposta fornita da Northpointe alle accuse di pregiudizi razziali nelle scale di rischio COMPAS mosse da ProPublica. La compagnia, entrata a far parte recentemente della più grande equivant Inc., ha dichiarato: "la modellazione predittiva è un campo specializzato all'interno della statistica e l'uso appropriato e l'interpretazione di modelli predittivi validi richiedono una solida comprensione delle tecniche e delle sfumature metodologiche comuni a questo tipo di lavoro. Il nostro esame dettagliato di come ProPublica ha condotto la sua analisi ha rivelato diversi errori statistici e tecnici, come modelli di regressione mal specificati, termini di classificazione e misure di discriminazione definiti in modo errato, l'interpretazione e l'uso non corretto degli errori del modello e altro ancora. Questi errori hanno portato a una falsa conclusione di pregiudizio razziale; non crediamo che le conclusioni tratte siano

in realtà supportate in alcun modo dai campioni di dati utilizzati" [16].
L'azienda ha fornito una risposta formale all'articolo di ProPublica attraverso il report "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity" [17]. Questa relazione tecnica presenta i risultati tecnici dopo un'attenta revisione della metodologia statistica utilizzata da ProPublica e i risultati dell'analisi approfondita degli stessi dati. Il documento contiene una spiegazione degli errori identificati dalla loro revisione e una discussione tecnica approfondita su tre risultati chiave della loro analisi metodologicamente corretta dei dati:

- ProPublica si è concentrata su statistiche di classificazione che non tenevano conto dei diversi tassi base di recidiva per i neri e i bianchi. L'uso di queste statistiche ha portato ad affermazioni false nel loro articolo, che sono state ripetute successivamente in interviste e articoli sui media nazionali.

- Se si utilizzano le statistiche di classificazione corrette, i dati non confermano l'affermazione di ProPublica di pregiudizio razziale nei confronti dei neri.

- L'interpretazione dei risultati nei campioni utilizzati da ProPublica dimostra che la General Recidivism Risk Scale (GRRS) e la Violent Recidivism Risk Scale (VRRS) sono ugualmente accurate per neri e bianchi.

Queste sono le affermazioni da parte di Northpointe riguardo all'articolo realizzato da ProPublica rispetto il loro prodotto.
A distanza di sei mesi dall'accaduto, Northpointe Inc. si è fusa con altre due società la Courtview Justice Solutions Inc. e la Constellation Justice Systems Inc., sotto il nome di "Equivant". L'annuncio della fusione confermava anche che "tutte le linee di prodotti esistenti sarebbero rimaste intatte" (Equivant, 2017). Si può quindi ipotizzare che Equivant abbia continuato a calcolare le valutazioni del rischio più o meno nello stesso modo in cui lo aveva fatto Northpointe e che abbia proseguito sulla stessa linea.
Questo intenso dibattito pubblico tra Northpointe e ProPublica ha inevitabilmente acceso la discussione tra gli scienzati, gli esperti della comunità scientifica e il pubblico in generale. Stabilire con certezza chi abbia ragione non è facile, a questo proposito sono stati pubblicati numerosi articoli nella letteratura a sostegno dell'una o dell'altra parte, tuttavia rimane ancora una questione aperta.

Capitolo 4

Riproduzione delle analisi di ProPublica

4.1 Analisi

In questa sezione mi soffermerò su come ho riprodotto le analisi effettuate da ProPublica e commenterò i risultati ottenuti verificando se sono in linea con quelli riportati nel loro studio [18]. Innanzitutto, ho cominciato con l'effettuare la raccolta dei dati che ho organizzato in data frames per facilitarne la consultazione, successivamente ho proseguito con il filtraggio a causa di alcuni dati mancanti.

Ho deciso di rimuovere i dati che non soddisfavano i seguenti requisiti:

- La data di imputazione di un reato valutato da COMPAS non era entro 30 giorni dal momento dell'arresto, ho ritenuto che per motivi di qualità dei dati non fosse corretto considerarli.
- In caso di assenza di un caso COMPAS ho codificato il flag di recidiva is_recid con il valore -1.
- In modo analogo, ho rimosso le infrazioni ordinarie al codice della strada, quelle con un c_charge_degree di '0', che non comportano la reclusione in carcere.
- Inoltre ho incluso soltanto le righe che rappresentano persone recidive nel corso di due anni o che hanno trascorso almeno due anni fuori da un istituto penitenziario.

4.1.1 Classificazione del rischio di recidiva

Dopo aver filtrato tutti i dati, ho eseguito le prime manipolazioni per valutare il rischio di recidiva. Inizialmente l'analisi ha preso in considerazione la semplice distribuzione dei punteggi COMPAS in decili tra bianchi e neri. Di seguito, ho riportato la distribuzione di questi punteggi per 6172 imputati che non erano stati arrestati per un nuovo

reato o che erano recidivi entro due anni.

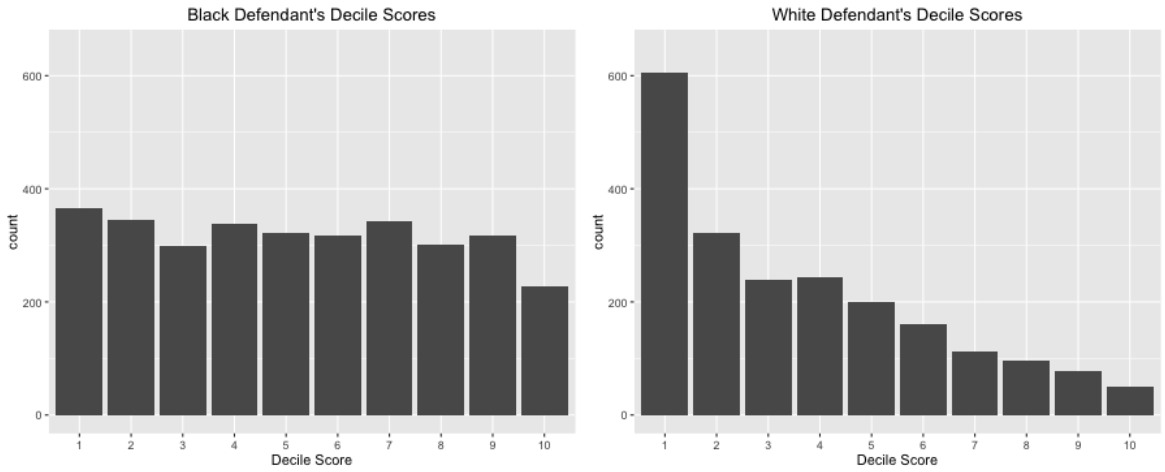

Figura 4.1

Questi istogrammi mostrano che i punteggi degli imputati bianchi si concentrano verso le categorie a rischio più basso, mentre gli imputati neri sono stati distribuiti in modo uniforme tra i punteggi. Il campione di due anni comprende 6172 imputati così distribuiti: African-American 3175, Asian 31, Caucasian 2103, Hispanic 509, Native American 11, Other 343, di cui 1.175 imputati donne e 4.997 imputati uomini.
In questo campione, gli imputati recidivi entro i due anni sono stati 2.809.

4.1.2 Classificazione del rischio di recidiva violenta

Gli istogrammi del punteggio di rischio di recidiva violenta di COMPAS mostrano una disparità nella distribuzione del punteggio tra imputati bianchi e neri. Il campione che ho utilizzato per testare il punteggio di recidiva violenta di COMPAS era leggermente più piccolo rispetto al punteggio di recidiva generale: 4.020 imputati, 1.918 imputati neri e 1.459 imputati bianchi. I recidivi violenti erano 652.

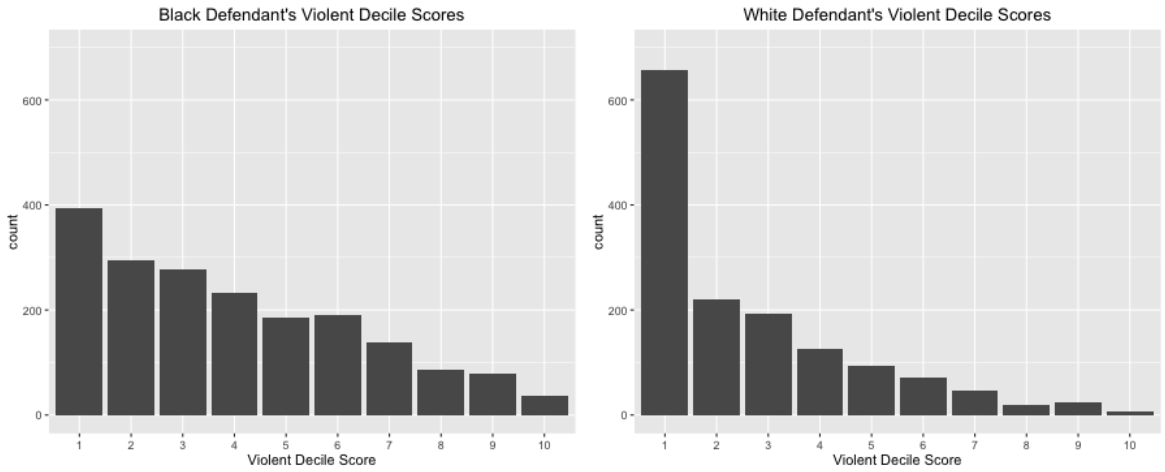

Figura 4.2

Sebbene esista una chiara differenza tra le distribuzioni dei punteggi COMPAS per gli imputati bianchi e neri, il semplice esame delle distribuzioni non tiene conto di altri fattori demografici e comportamentali. Per verificare la presenza di disparità razziali nel punteggio prendendo in considerazione anche altri fattori, ho utilizzato un modello di regressione logistica che considerava l'etnia, l'età, i precedenti penali, la recidiva futura, il grado di accusa, il genere e l'età.

4.1.3 Modello di regressione logistica di recidiva

In questa sezione, ho usato i fattori citati poco fa per modellare le probabilità di ottenere un punteggio COMPAS più alto. Il modello logistico ha rilevato che il fattore più predittivo di un punteggio di rischio più elevato è l'età. Gli imputati di età inferiore ai 25 anni hanno una probabilità 2,5 volte maggiore di ottenere un punteggio più alto rispetto agli imputati di mezza età, anche quando si controllano i reati precedenti, la criminalità futura, l'etnia e il genere. Anche l'etnia è abbastanza predittiva di un punteggio più alto. Sebbene gli imputati neri avessero tassi di recidiva complessivamente più elevati, una volta aggiustati per questa differenza e altri fattori, avevano il 45% di probabilità in più di ottenere un punteggio più alto rispetto ai bianchi.
Sorprendentemente, dati i loro livelli di criminalità complessivamente più bassi, le imputate donne hanno il 19.4% di probabilità in più di ottenere un punteggio più alto rispetto agli uomini, controllando gli stessi fattori.

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402

*Note: *p<0.1; **p<0.05; ***p<0.01*

Figura 4.3

4.1.4 Modello di regressione logistica di recidiva violenta

Per quanto riguarda il rischio di recidiva violenta, ho analizzato 4.020 persone a cui è stato assegnato un punteggio per la recidiva violenta in un periodo di due anni (escluso il tempo trascorso in carcere). In seguito, ho eseguito un modello di regressione simile a quello appena presentato.
L'età è risultata un fattore predittivo ancora più forte di un punteggio più alto di recidiva violenta. La regressione ha mostrato che i giovani imputati avevano una probabilità 6.4 volte maggiore di ottenere un punteggio più alto rispetto agli imputati di mezza età, quando si correggevano per i precedenti penali, il genere, l'etnia e la futura recidiva violenta. Anche l'etnia era predittiva di un punteggio più alto per la recidiva violenta. Gli imputati neri hanno il 77.3% di probabilità in più rispetto agli imputati bianchi di ricevere un punteggio più alto, correggendo per i precedenti penali e la futura recidiva violenta.

Risk of Violent Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	-0.729*** (0.127)
Age: Greater than 45	-1.742*** (0.184)
Age: Less than 25	3.146*** (0.115)
Black	0.659*** (0.108)
Asian	-0.985 (0.705)
Hispanic	-0.064 (0.191)
Native American	0.448 (1.035)
Other	-0.205 (0.225)
Number of Priors	0.138*** (0.012)
Misdemeanor	-0.164* (0.098)
Two Year Recidivism	0.934*** (0.115)
Constant	-2.243*** (0.113)
Observations	4,020
Akaike Inf. Crit.	3,022.779

*Note: *p<0.1; **p<0.05; ***p<0.01*

Figura 4.4

4.1.5 Modello di regressione di Cox

Per verificare l'accuratezza predittiva complessiva di COMPAS, ho applicato ai dati un modello di Cox proportional hazards, la stessa tecnica utilizzata da Northpointe nel suo studio di validazione. Il modello di Cox permette di confrontare i tassi di recidiva controllando il tempo. Poiché non controllo altri fattori, come la criminalità dell'imputato, è possibile includere più persone nel modello di Cox. Per questa analisi

la dimensione del campione era di 10.314 imputati (3.569 imputati bianchi e 5.147 imputati neri).
Ho considerato le persone del set di dati come "a rischio" dal giorno in cui hanno ricevuto il punteggio COMPAS fino al giorno in cui hanno commesso un nuovo reato o al 1 aprile 2016, a seconda di quale dei due eventi si è verificato per primo. Ho deciso di rimuovere le persone dal set di rischio durante la loro permanenza in carcere. La variabile indipendente nel modello di Cox era il punteggio di rischio categorico COMPAS.
Il modello di Cox ha mostrato che le persone con punteggi elevati avevano una probabilità di recidiva 3.5 volte superiore rispetto a quelle della categoria bassa (punteggi da 1 a 4). Lo studio di Northpointe ha rilevato che le persone con punteggi elevati (da 8 a 10) avevano una probabilità di recidiva 5.6 volte superiore. Entrambi i risultati indicano che il punteggio ha un valore predittivo.
Un diagramma di sopravvivenza Kaplan Meier mostra anche una chiara differenza nei tassi di recidiva tra ogni livello di punteggio COMPAS.

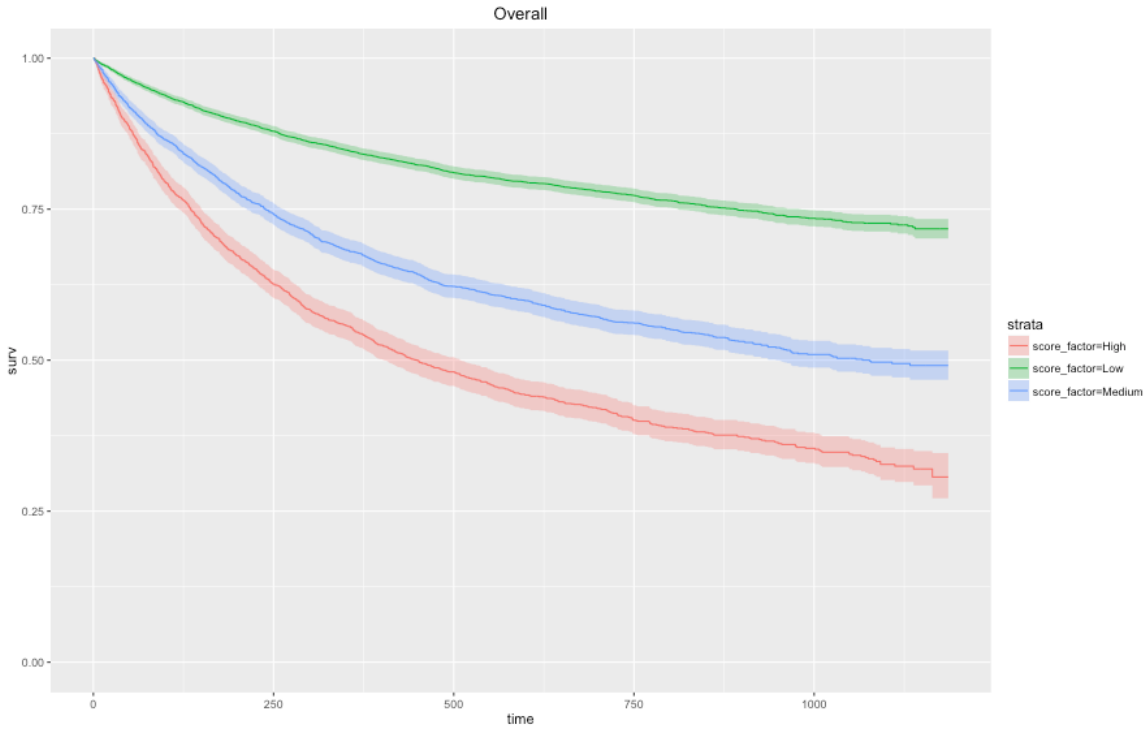

Figura 4.5

Nel complesso, la regressione di Cox ha ottenuto un punteggio di concordanza del 63.6%. Ciò significa che per qualsiasi coppia di imputati selezionati a caso nel campione, il sistema COMPAS è in grado di classificare accuratamente il loro rischio di recidiva nel 63.6% dei casi (ad esempio, se una persona della coppia è recidiva, quella coppia conterà come una corrispondenza riuscita se anche quella persona aveva un punteggio più alto). Nel suo studio, Northpointe ha riportato una concordanza leggermente superiore: 68%.
Eseguendo il modello di Cox sui punteggi di rischio sottostanti, classificati da 1 a 10, piuttosto che sugli intervalli basso, medio e alto, si è ottenuta una concordanza

leggermente superiore, pari al 66.4%. Entrambi i risultati sono inferiori a quella che Northpointe definisce una soglia di affidabilità. "Secondo diversi articoli recenti, una regola empirica è che AUC pari o superiori a .70 indicano in genere un'accuratezza predittiva soddisfacente, mentre misure comprese tra .60 e .70 suggeriscono un'accuratezza predittiva da bassa a moderata", afferma l'azienda nel suo studio.
Il punteggio di recidiva violenta di COMPAS aveva una concordanza del 65.1%.
Il sistema COMPAS prevede la recidiva in modo disomogeneo tra i generi. Secondo le stime di Kaplan-Meier, le donne classificate ad alto rischio hanno recidivato a un tasso del 47.5% nei due anni successivi al punteggio. Tuttavia, gli uomini classificati ad alto rischio sono recidivi ad un tasso molto più alto, pari al 61.2%, nello stesso periodo di tempo. Ciò significa che una donna ad alto rischio ha un rischio di recidiva molto più basso di un uomo ad alto rischio, un fatto che può essere trascurato dai funzionari delle forze dell'ordine che interpretano il punteggio.

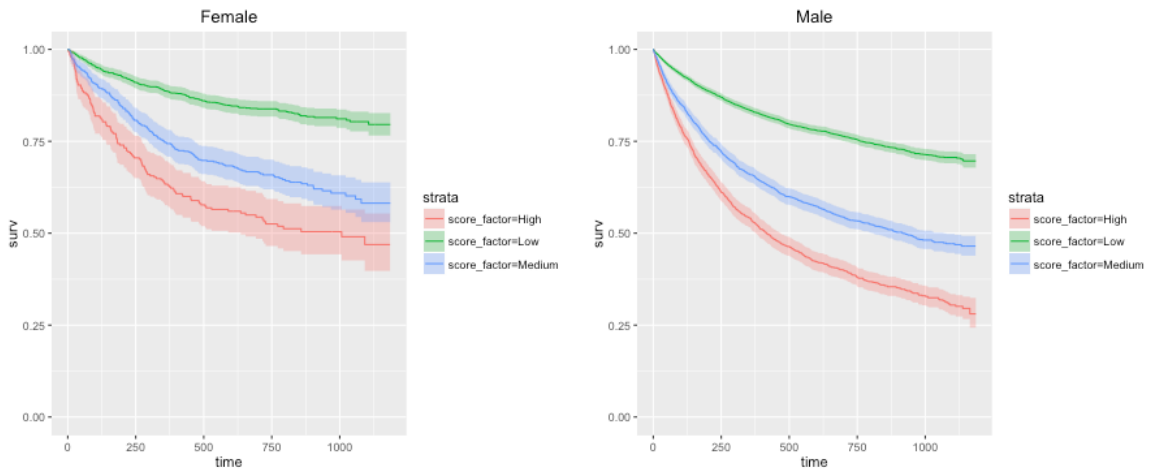

Figura 4.6

4.1.6 Osservazioni rispetto ai risultati ottenuti

Durante le analisi ho osservato che l'accuratezza predittiva del punteggio di recidiva COMPAS era coerente tra le diverse etnie: 62.5% per gli imputati bianchi e 62.3% per quelli neri. Gli autori dello studio Northpointe hanno riscontrato una piccola differenza nei punteggi di concordanza per etnia: 69% per gli imputati bianchi e 67% per quelli neri. In tutte le categorie di rischio, gli imputati neri commettono recidiva con tassi più elevati.
In seguito, ho aggiunto un termine di interazione etnia-punteggio al modello di Cox. Questo termine mi ha permesso di valutare se la differenza di recidiva tra un punteggio alto e uno basso fosse diversa per gli imputati neri e per quelli bianchi.
Il coefficiente sui punteggi elevati per gli imputati neri è quasi statisticamente significativo (0.0574). Gli imputati bianchi ad alto rischio hanno una probabilità di recidiva 3.61 volte superiore a quella degli imputati bianchi a basso rischio, mentre gli imputati

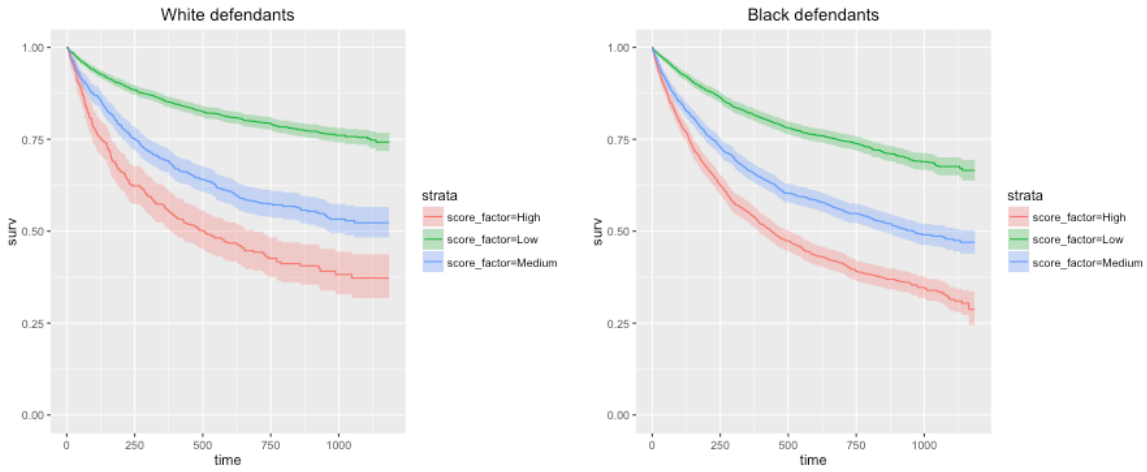

Figura 4.7

neri ad alto rischio hanno una probabilità di recidiva solo 2.99 volte superiore a quella degli imputati neri a basso rischio. Anche gli hazard ratio per gli imputati a medio rischio rispetto a quelli a basso rischio sono diversi a seconda della etnia: 2.32 per gli imputati bianchi e 1.95 per quelli neri. A causa del divario tra gli hazard ratio, è possibile concludere che il punteggio funziona in modo diverso tra i sottogruppi etnici.

Risk of General Recidivism Cox Model (with Interaction Term)	
Black	0.279*** (0.061)
Asian	-0.777 (0.502)
Hispanic	-0.064 (0.097)
Native American	-1.255 (1.001)
Other	0.014 (0.110)
High Score	1.284*** (0.084)
Medium Score	0.843*** (0.071)
Black:High	-0.190* (.100, p: 0.0574)
Asian:High	1.316* (0.768)
Hispanic:High	-0.119 (0.198)
Native American:High	1.956* (.083)
Other:High	0.415 (0.259)
Black:Medium	-0.173* (.091, p: 0.0578)
Asian:Medium	0.986 (0.711)
Hispanic:Medium	0.065 (0.164)
Native American:Medium	1.390 (1.120)
Other:Medium	-0.334 (0.232)
Observations	13,344
R2	0.072
Max. Possible R2	0.990
Log Likelihood	-30,280.410
Wald Test	988.830*** (df = 17)
LR Test	993.709*** (df = 17)
Score (Logrank) Test	1,104.894*** (df = 17)
*Note: *p<0.1; **p<0.05; ***p<0.01*	

Figura 4.8

Successivamente ho condotto un'analisi simile sul punteggio di recidiva violenta di COMPAS, ma non ho trovato un risultato simile. In questo caso, ho riscontrato che il termine di interazione tra etnia e punteggio non era significativo, il che significa che non c'è una differenza significativa tra i rischi degli imputati neri ad alto e basso rischio e quelli bianchi ad alto e basso rischio.

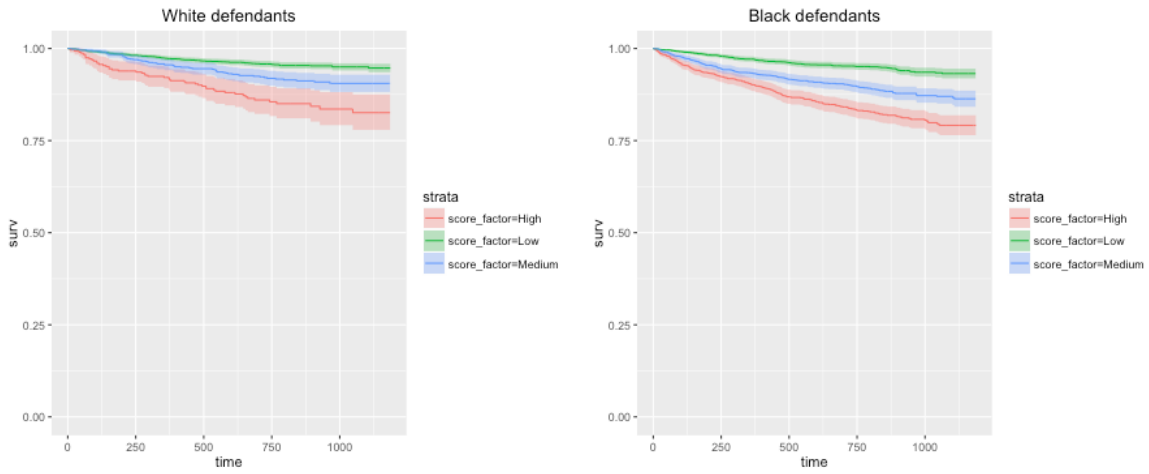

Figura 4.9

4.1.7 Valutazione dei dati in tabelle di contingenza

Infine, ho verificato se alcuni tipi di errori, falsi positivi e falsi negativi, fossero distribuiti in modo disomogeneo tra le etnie. Per fare questo ho usato delle tabelle di contingenza per determinare i tassi relativi, seguendo l'analisi delineata nel documento del 2006 della Salvation Army. Ho rimosso dal set di dati le persone per le quali avevo meno di due anni di informazioni sulla recidiva. La popolazione rimanente era di 7.214 persone, un po' più grande del campione dei modelli logistici di cui sopra, perché per questa analisi non c'era bisogno delle informazioni sul caso dell'imputato. Come nell'analisi di regressione logistica, ho contrassegnato i punteggi diversi da "basso" come rischio più elevato. La Figura 5.8, riportata nella pagina successiva, mostra l'andamento del punteggio di recidiva COMPAS.

All Defendants			Black Defendants			White Defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		
PPV: 0.61			PPV: 0.63			PPV: 0.59		
NPV: 0.69			NPV: 0.65			NPV: 0.71		
LR+: 1.94			LR+: 1.61			LR+: 2.23		
LR-: 0.55			LR-: 0.51			LR-: 0.62		

Figura 4.10

Queste tabelle di contingenza rivelano che è più probabile che l'algoritmo classifichi erroneamente un imputato nero come ad alto rischio rispetto a un imputato bianco.

Gli imputati neri che non commettono recidiva hanno quasi il doppio delle probabilità di essere classificati da COMPAS come ad alto rischio rispetto alle loro controparti bianche (45% contro 23%). Tuttavia, gli imputati neri che hanno ottenuto un punteggio più alto sono recidivi un po' più spesso degli imputati bianchi (63% contro 59%). Il test tende a commettere l'errore opposto con i bianchi, nel senso che è più probabile che preveda erroneamente che i bianchi non commettano altri reati se rilasciati rispetto agli imputati neri. Stando a quanto appena affermato, COMPAS ha sottoclassificato i recidivi bianchi come a basso rischio il 70, 5% più spesso di quelli neri (48% contro 28%). Infatti l'indice di probabilità per gli imputati bianchi era leggermente più alto, 2.23, rispetto a quello per gli imputati neri, 1.61.
Ho anche verificato se restringere la definizione di rischio elevato includendo solo il punteggio alto di COMPAS, anziché includere sia il punteggio medio che quello alto, cambiasse i risultati della nostra analisi. In questo scenario, gli imputati neri avevano una probabilità tre volte superiore a quella degli imputati bianchi di essere ingiustamente classificati ad alto rischio (16% contro 5%).
In particolare, ho trovato risultati simili per il punteggio di recidiva violenta di COMPAS. Come in precedenza, ho calcolato delle tabelle di contingenza in base all'andamento del punteggio che ho riportato nella Figura 5.10.

All Defendants		
	Low	High
Survived	4121	1597
Recidivated	347	389
FP rate: 27.93		
FN rate: 47.15		
PPV: 0.20		
NPV: 0.92		
LR+: 1.89		
LR-: 0.65		

Black defendants		
	Low	High
Survived	1692	1043
Recidivated	170	273
FP rate: 38.14		
FN rate: 38.37		
PPV: 0.21		
NPV: 0.91		
LR+: 1.62		
LR-: 0.62		

White defendants		
	Low	High
Survived	1679	380
Recidivated	129	77
FP rate: 18.46		
FN rate: 62.62		
PPV: 0.17		
NPV: 0.93		
LR+: 2.03		
LR-: 0.77		

Figura 4.11

Infine, ho osservato che gli imputati neri hanno una probabilità doppia rispetto a quelli bianchi di essere classificati erroneamente come ad alto rischio di recidiva violenta, mentre i recidivi bianchi sono stati classificati erroneamente come a basso rischio il 63.2% più spesso degli imputati neri. Gli imputati neri classificati come a più alto rischio di recidiva violenta hanno commesso recidiva a un tasso leggermente superiore rispetto agli imputati bianchi (21% contro 17%), e il rapporto di probabilità per gli imputati bianchi era più alto, 2.03, rispetto a quello per gli imputati neri, 1.62.

4.2 Analisi intersezionale

In questa sezione, ho ampliato le analisi condotte da ProPublica effettuando un'analisi intersezionale sulla base degli stessi dati utilizzati in precedenza. Si tratta di uno strumento d'analisi per studiare, comprendere e affrontare i modi in cui il sesso e il genere

si intersecano con altre caratteristiche/identità personali e il modo in cui tali intersezioni contribuiscono insieme ad un'unica esperienza di discriminazione. Parte dalla premessa che le persone hanno identità multiple e stratificate derivanti dalle relazioni sociali, dalla storia e dall'operato delle strutture di potere. L'analisi intersezionale mira a rivelare tali identità multiple mettendo in luce i diversi tipi di discriminazione intersezionale e discriminazione multipla, nonché gli svantaggi conseguenti alla combinazione delle diverse identità e all'intersezione di sesso e genere con altri motivi [19].

Race	Metrics	Male	Female
Caucasian	FP rate	21.25	30.16
Caucasian	FN rate	48.89	43.22
African-American	FP rate	46.12	40.49
African-American	FN rate	27.69	29.96

Intersectional analysis

Tabella 4.1

Nella Tabella 5.1, ho riportato il tasso di Falsi Positivi (FP) e Falsi Negativi (FN) in base all'etnia e al sesso di appartenenza. In particolare, se confrontiamo i valori ottenuti per gli uomini si nota che la percentuale di falsi positivi registrata per gli individui afroamericani (46.12) è più del doppio rispetto a quella degli uomini di origini caucasiche (21.25). Ciò significa che l'algoritmo COMPAS tende a predire erroneamente il rischio di recidiva per gli afroamericani, il 2.17 delle volte in più rispetto agli individui caucasici. Nel caso delle donne, i risultati registrati sono più rassicuranti anche se emerge comunque la presenza di una disparità pari all'1.34. Infatti alle donne di origine afroamericana viene attribuito un punteggio di rischio errato con un tasso del 40.49, mentre alle donne di origine caucasica questo avviene con un 30.16.
Per quanto riguarda il tasso di falsi negativi, si osserva la situazione analoga che vede gli individui di origini caucasiche favoriti rispetto alla controparte. Più interessante è invece l'analisi tra individui appartenenti alla stessa classe di provenienza. Si noti infatti che all'interno dello stesso gruppo il tasso varia anche a seconda del sesso degli individui considerati. Le donne risultano le più penalizzate se si considera l'etnia caucasica, mentre vale il contrario per gli afroamericani che vede gli uomini essere sottoposti a giudizi più severi.
Nella Tabella 5.2, ho inserito altre metriche significative quali: PPV (Positive Predictive Value), NPV (Negative Predictive Value), LR+ (Positive Likelihood Ratio) e LR- (Negative Likelihood Ratio). Tali metriche riflettono a loro volta quanto già affermato in precedenza. In particolare, la PPV, o precisione, definita come:

$$PPV = \frac{TP}{TP + FP} \tag{4.1}$$

rappresenta la probabilità che una persona sia realmente recidiva in relazione a quelle classificate come recidive. In altre parole, è la percentuale di individui recidivi fra quelli

Race	Metrics	Male	Female
Caucasian	PPV	0.62	0.50
Caucasian	NPV	0.70	0.75
Caucasian	LR+	2.41	1.88
Caucasian	LR-	0.62	0.62
African-American	PPV	0.65	0.51
African-American	NPV	0.62	0.77
African-American	LR+	1.57	1.73
African-American	LR-	0.51	0.50

Intersectional analysis

Tabella 4.2

identificati dall'algoritmo. La NPV definita come:

$$NPV = \frac{TN}{TN + FN} \tag{4.2}$$

rappresenta la probabilità che una persona sia non recidiva dato che non è stata classificata come tale. In altre parole, è la percentuale di individui non recidivi fra quelli identificati dall'algoritmo. La LR+ definita come:

$$LR+ = \frac{TPR}{FPR} \tag{4.3}$$

che può essere espressa anche in funzione della sensitivity (o True Positive Rate) e della specificity (o True Negative Rate) nella forma:

$$LR+ = \frac{sensitivity}{1 - specificity} \tag{4.4}$$

corrisponde alla probabilità che un individuo sia recidivo dato che gli è stato assegnato dall'algoritmo un punteggio elevato diviso la probabilità che lo stesso individuo non sia recidivo dato che gli è stato assegnato un punteggio elevato. In modo analogo, esiste anche la metrica LR- che considera i casi negativi:

$$LR- = \frac{FNR}{TNR} \tag{4.5}$$

che può essere espressa anche in funzione della sensitivity e della specificity nella forma:

$$LR- = \frac{1 - sensitivity}{specificity} \tag{4.6}$$

corrisponde alla probabilità che un individuo sia recidivo dato che gli è stato assegnato dall'algoritmo un punteggio basso diviso la probabilità che lo stesso individuo non sia recidivo dato che gli è stato assegnato un punteggio basso.
In seguito, ho considerato ulteriori metriche nell'analisi intersezionale quali: il FDR (False Discovery Rate) e il FOR (False Omission Rate).

Race	Metrics	Male	Female
Caucasian	FDR	0.38	0.50
Caucasian	FOR	0.30	0.25
African-American	FDR	0.35	0.49
African-American	FOR	0.38	0.23

Intersectional analysis

Tabella 4.3

Il FDR, la cui formula è data da:

$$FDR = 1 - PPV = 1 - \frac{TP}{TP + FP} = \frac{FP}{TP + FP} \tag{4.7}$$

rappresenta il complementare della PPV, ovvero la probabilità di venire erroneamente classificati come recidivi. Mentre il FOR, la cui definizione è data da:

$$FOR = 1 - NPV = 1 - \frac{TN}{TN + FN} = \frac{FN}{TN + FN} \tag{4.8}$$

rappresenta il complementare della NPV, ovvero la probabilità di venire erroneamente classificati come non recidivi. Le metriche FDR e FOR producono risultati che dovrebbero collocarsi in un range compreso tra 0 e 1, dove 0 rappresenta il valore migliore e 1 il valore peggiore. Osservando i valori raccolti nella Tabella 4.3, non emergono particolari disparità tra i due gruppi considerati, Caucasian ed African-American. Tuttavia, i valori riportati sono abbastanza elevati ed indicano un'importante tasso di errore. Ad esempio, nel caso di Caucasian female si osserva un FDR pari a 0.50, ciò significa che l'algoritmo COMPAS tende a predire correttamente la recidiva nel 50% dei casi, in altre parole il livello di precisione è pari a quello che si avrebbe lanciando una moneta. Infine, nelle tabelle che seguono ho riportato il campione di individui su cui sono state condotte le analisi.

Caucasian men	Low	High	Percentage
Survived	882	238	0.59
Recidivated	375	392	0.41
Total: 1887.00			

Caucasian

Tabella 4.4

Caucasian women	Low	High	Percentage
Survived	257	111	0.65
Recidivated	86	113	0.35
Total: 567.00			

Caucasian

Tabella 4.5

African-American men	Low	High	Percentage
Survived	749	641	0.46
Recidivated	458	1196	0.54
Total: 3044.00			

African-American

Tabella 4.6

African-American women	Low	High	Percentage
Survived	241	164	0.62
Recidivated	74	173	0.38
Total: 652.00			

African-American

Tabella 4.7

Capitolo 5

Aequitas - Bias and Fairness Audit

5.1 Caratteristiche e funzionalità

Aequitas è in grado di verificare i sistemi di intelligenza artificiale per individuare azioni o risultati distorti che si basano su ipotesi false o distorte su vari gruppi demografici. Utilizzando una libreria Python e un'interfaccia a riga di comando, è possibile caricare semplicemente i dati del sistema da verificare, configurare le metriche di distorsione per i gruppi di attributi protetti di interesse e per i gruppi di riferimento, quindi lo strumento genera rapporti sulle distorsioni [20].

5.1.1 Misurare i bias e la fairness

Un gran numero di problemi di politica pubblica e di utilità sociale, in cui i sistemi di IA vengono utilizzati per aiutare gli esperti umani a prendere decisioni, condividono alcune caratteristiche comuni, tra cui una distribuzione obliqua delle classi e la metrica di interesse è la precisione al massimo k, al contrario dell'accuratezza di molti altri problemi di apprendimento automatico. Questo perché spesso le nostre risorse di intervento sono limitate e possiamo intervenire solo su un numero ridotto (k) di entità. L'obiettivo del modello ML è quello di stabilire con precisione la priorità delle k migliori, il che equivale a massimizzare la precisione delle k migliori. La Figura 7.1 illustra una linea temporale comune dei sistemi di IA per la politica e il bene sociale. In base al punteggio di rischio (previsione) prodotto dal modello ML, le entità vengono classificate e, spesso con l'intervento di un esperto umano, vengono selezionate le prime k entità su cui intervenire. Questi interventi possono essere di tipo assistenziale (aiutare gli individui a ricevere assistenza abitativa per ridurre il rischio di rimanere in futuro senza casa) o talvolta punitivo (ispezioni degli alloggi che comportano multe e costi di riparazione in caso di violazioni).
Un compito tradizionale di classificazione binaria che utilizza l'apprendimento supervisionato consiste nell'apprendimento di un predittore $\hat{Y} \in \{0, 1\}$, che mira a prevedere il vero risultato $Y \in \{0, 1\}$ di un dato punto di dati dall'insieme di caratteristiche X,

Figura 5.1: Linea di decisione algoritmica per i problemi di politica pubblica e di interesse sociale.

sulla base di dati di addestramento etichettati. Molti problemi di politica pubblica possono essere formulati come problemi di valutazione statistica del rischio, in cui si assegna un punteggio reale $S \in [0, 1]$ a ogni entità (punto dati) e si prende una decisione $\hat{Y}$ in base al punteggio, tipicamente selezionando un numero predefinito (k) di entità che dovrebbero essere classificate come positive. Dopo aver ordinato le entità in base a S, il predittore binario è definito come $\hat{Y} = 1$ se $R \geq sk$ dove sk è il punteggio della k-esima entità ordinata. Le definizioni principali di questa sottosezione sono le seguenti [20]:

- **Punteggio** - $S \in [0, 1]$ è un punteggio a valore reale assegnato a ciascuna entità dal predittore.
- **Decisione** - $\hat{Y} \in \{0, 1\}$ è una previsione binaria assegnata a una data entità (punto dati), basata su una soglia del punteggio (ad esempio, top K).
- **Risultato vero** - $Y \in \{0, 1\}$ è la vera etichetta binaria di una data entità.

5.1.2 Definizione dei gruppi

Si consideri un attributo che può assumere più valori $A = \{a_i, a_2, ..., a_n\}$ che può essere o meno un sottoinsieme di X, ad esempio $gender = \{femmina, maschio, altro\}$. Definiamo un gruppo $g(ai)$ come un insieme di entità (punti dati) che hanno in comune uno specifico valore dell'attributo $A = a_i$, per esempio $gender = female$ che corrisponde a tutte le femmine del dataset. Dati tutti i gruppi definiti dall'attributo A, le previsioni $\hat{Y}$ e il risultato vero Y per ogni entità di ciascun gruppo, è possibile ora discutere le metriche di gruppo. Le definizioni principali sulla definizione dei gruppi per la valutazione dei bias e dell'equità sono le seguenti [20]:

- **Attributo** - $A = \{a_i, a_2, ..., a_n\}$ è un attributo con più valori, ad esempio, $genere = \{femmina, maschio, altro\}$.
- **Gruppo** - $g(ai)$ è un gruppo di tutte le entità che condividono lo stesso valore dell'attributo, ad esempio, $gender = femmina$.

- **Gruppo di riferimento** - $g(ar)$ è uno dei gruppi di A che viene utilizzato come riferimento per calcolare le misure di bias.
- **Labeled Positive** - LP_g è il numero di entità etichettate come positive all'interno di un gruppo.
- **Labeled Negative** - LN_g è il numero di entità etichettate come negative all'interno di un gruppo.
- **Prevalenza** - $Prev_g = LP_g/|g| = Pr(Y = 1|A = a_i)$ è la frazione di entità all'interno di un gruppo il cui esito vero era positivo.

5.1.3 Metriche del gruppo di distribuzione

Pertanto ora è possibile definire le metriche decisionali a livello di gruppo. Utilizzo due metriche (Predicted Prevalence e Predicted Positive Rate) che si preoccupano solo della distribuzione delle entità tra i gruppi nel set selezionato per l'intervento (top k) e quindi non utilizzano i risultati reali (etichette). Definiamo le metriche distributive dei gruppi come segue [20]:

- **Predicted Positive** - PP_g è il numero di entità all'interno di un gruppo in cui la decisione è positiva, cioè, $\hat{Y} = 1$.
- **Total Predictive Positive** - $K = \sum_{A=a_1}^{A=a_n} PP_{g(a_i)}$ è il numero totale di entità previste positive nei gruppi definiti da A.
- **Predicted Negative** - PN_g è il numero di entità all'interno di un gruppo la cui decisione è negativa, cioè $\hat{Y} = 0$.
- **Predicted Prevalence** - $PPrev_g = PP_g/|g| = Pr(\hat{Y} = 1|A = a_i)$ è la frazione di entità all'interno di un gruppo che sono state predette come positive.
- **Predicted Positive Rate** - $PPR_g = PP_g/K = Pr(a = a_i|\hat{Y} = 1)$ è la frazione delle entità previste come positive che appartengono a un certo gruppo.

5.1.4 Metriche di gruppo basate su errori

In questa sezione definiamo le metriche di gruppo che richiedono il calcolo del risultato vero (etichetta). Ci concentriamo sugli errori di tipo I (falsi positivi) e II (falsi negativi) nei diversi gruppi. Nel contesto delle politiche pubbliche e del bene sociale, l'obiettivo è evitare errori sproporzionati in gruppi specifici. Utilizziamo quattro diverse metriche di gruppo basate sugli errori, definite come segue [20]:

- **False Positive** - FP_g è il numero di entità del gruppo con $\hat{Y} = 1 \wedge Y = 0$.
- **False Negative** - FN_g è il numero di entità del gruppo con $\hat{Y} = 0 \wedge Y = 1$.

- **True Positive** - TP_g è il numero di entità del gruppo con $\hat{Y} = 1 \wedge Y = 1$.
- **True Negative** - TN_g è il numero di entità del gruppo con $\hat{Y} = 0 \wedge Y = 0$.
- **False Discovery Rate** - $FDR_g = FP_g/PP_g = \frac{Pr(Y=0 \wedge \hat{Y}=1)}{Pr(\hat{Y}=1)}$ è la frazione di falsi positivi di un gruppo all'interno del gruppo di positivi previsti nel gruppo.
- **False Omission Rate** - $OR_g = FN_g/PN_g = \frac{Pr(Y=1 \wedge \hat{Y}=0)}{Pr(\hat{Y}=0)}$ è la frazione di falsi negativi di un gruppo all'interno del gruppo negativo previsto.
- **False Positive Rate** - $FPR_g = FP_g/LN_g = Pr(\hat{Y} = 1|Y = 0, A = a_i)$ è la frazione di falsi positivi di un gruppo all'interno del negativo etichettato del gruppo stesso.
- **False Negative Rate** - $FNR_g = FN_g/LP_g = Pr(\hat{Y} = 0|Y = 1, A = a_i)$ è la frazione di falsi negativi di un gruppo all'interno dei positivi etichettati del gruppo.

Capitolo 6

Analisi Aequitas di COMPAS

6.1 The Bias Report

Aequitas ha redatto il documento "The Bias Report" con l'obiettivo di raccogliere i risultati ottenuti dalla loro analisi e fare chiarezza sul processo di individuazione di eventuali pregiudizi.
Nel fare questo ha seguito un processo suddiviso in diversi steps [21]:

1. **Caricamento dei dati:** Innanzitutto hanno caricato il set di dati puliti/originale, disponibile nella pagina GitHub di ProPublica [22].

2. **Selezione dei gruppi protetti:** In seguito al dibattito tra ProPublica e Northpointe, si sono concentrati sull'etnia. Hanno selezionato un gruppo di riferimento personalizzato e hanno utilizzato l'etnia caucasica come gruppo di riferimento. Le loro metriche riflettono quindi la fairness in relazione al gruppo storicamente dominante.

3. **Selezione delle metriche sulla fairness:** Successivamente, rifacendosi sempre al dibattito hanno selezionato i tassi di falsi positivi (False Positive Rates), i tassi di falsi negativi (False Negative Rates) e i tassi di falsa scoperta (False Discovery Rates).

4. **Scelta della soglia:** Infine, hanno deciso di mantenere il valore predefinito dell'80%. Ciò significa che qualsiasi metrica del gruppo compresa tra l'80% e il 125% della metrica del gruppo di riferimento è considerata equa e qualsiasi metrica al di fuori di questo intervallo è considerata ingiusta.

L'obiettivo dell'analisi era quello di individuare le seguenti metriche:

- **Parità del tasso di falsi positivi (False Positive Rate Parity)** - Assicura che tutti i gruppi protetti abbiano lo stesso tasso di falsi positivi del gruppo di riferimento.

- **Parità del tasso di falsa scoperta (False Discovery Rate Parity)** - Assicura che tutti i gruppi protetti abbiano una proporzione uguale di falsi positivi all'interno del set selezionato (rispetto al gruppo di riferimento).
- **Parità del tasso di falsi negativi (False Negative Rate Parity)** - Assicura che tutti i gruppi protetti abbiano lo stesso tasso di falsi negativi (rispetto al gruppo di riferimento).

6.2 Risultati ottenuti

In questa sezione ho riportato le tabelle che racchiudono i risultati raccolti per ogni metrica considerata.

Attribute Value	False Discovery Rate Parity	False Positive Rate Parity	False Negative Rate Parity
African-American	Passed	Failed	Failed
Asian	Failed	Failed	Failed
Caucasian	Ref	Ref	Ref
Hispanic	Passed	Passed	Passed
Native American	Failed	Failed	Failed
Other	Passed	Failed	Failed

Audit Results: Details by Protected Attributes

Tabella 6.1

Attribute Value	False Discovery Rate Disparity	False Positive Rate Disparity	False Negative Rate Disparity
African-American	0.91	1.91	0.59
Asian	0.61	0.37	0.7
Caucasian	1.0	1.0	1.0
Hispanic	1.12	0.92	1.17
Native American	0.61	1.6	0.21
Other	1.12	0.63	1.42

Audit Results: Bias Metrics Values

Tabella 6.2

Da queste tabelle si può notare come l'analisi rispetto alla metrica di False Positive Rate Parity non sia andata a buon fine, in quanto i valori ottenuti di False Positive Rate Disparity per la maggior parte dei gruppi non rientrano all'interno del range previsto $80\% - 125\%$. In particolare, rispetto al gruppo di riferimento "Caucasian" si osserva il seguente grado di disparità: 0.37 per il gruppo "Asian", 1.91 per il gruppo "African-American", 1.60 per il gruppo "Native American" e 0.63 per il gruppo "Other".

Risulta più soddisfacente l'analisi rispetto alla metrica di False Discovery Rate Parity con solo due gruppi che risultano non soddisfare i requisiti rispetto al gruppo di riferimento "Caucasian". Si tratta dei gruppi "Native American" e "Asian" con un tasso di disparità dello 0.61.

Attribute Value	Group Size Ratio	False Discovery Rate	False Positive Rate	False Negative Rate
African-American	0.51	0.37	0.45	0.28
Asian	0	0.25	0.09	0.33
Caucasian	0.34	0.41	0.23	0.48
Hispanic	0.09	0.46	0.21	0.56
Native American	0	0.25	0.38	0.1
Other	0.05	0.46	0.15	0.68

Audit Results: Group Metrics Values

Tabella 6.3

Al contrario, l'analisi della metrica False Negative Rate Parity mostra un alto grado di disparità dovuto ai valori decisamente più bassi di False Negative Rate rispetto al gruppo di riferimento "Caucasian". Infatti, il tasso di disparità per i vari gruppi risulta: lo 0.21 per il gruppo "Native American", lo 0.59 per il gruppo "African-American", lo 0.70 per il gruppo "Asian" e l'1.42 per il gruppo "Other".

Capitolo 7

Confronto risultati ProPublica ed Aequitas

7.1 Confronto dei valori delle metriche di gruppo

In questa sezione metterò a confronto i risultati che hanno prodotto le analisi condotte da ProPublica con quelli di Aequitas. In particolare, intendo verificare se i risultati ottenuti sono coerenti tra di loro o se presentano delle differenze significative.

All defendants	Low	High	Percentage
Survived	2681	1282	0.55
Recidivated	1216	2035	0.45
Total: 7214.00			

All defendants

Tabella 7.1: Analisi ProPublica

Nella Tabella 9.1, ho riportato il campione considerato nelle analisi di ProPublica estese a tutti gli imputati. Nella Tabella 9.2, sono raccolti i risultati riportati nelle varie metriche.

Metrics	Values
False positive rate	32.35
False negative rate	37.40
Specificity	0.68
Sensitivity	0.63
Prevalence	0.45
PPV	0.61
NPV	0.69
LR+	1.94
LR-	0.55

All defendants

Tabella 7.2: Risultati raccolti da ProPublica

Consideriamo ora il gruppo di riferimento "Caucasian". Confrontando i risultati della Tabella 9.4 con quelli della tabella 9.7, è possibile notare che i valori ricavati per le metriche False Positive Rate e False Negative Rate convergono, a meno di una piccola approssimazione, allo stesso valore. Infatti, la misura della prima metrica risulta pari circa al 23%, mentre la seconda assume un valore pari circa al 48%.

White defendants	Low	High	Percentage
Survived	1139	349	0.61
Recidivated	461	505	0.39
Total: 2454.00			

Caucasian

Tabella 7.3: Analisi ProPublica

Metrics	Values
False positive rate	23.45
False negative rate	47.72
Specificity	0.77
Sensitivity	0.52
Prevalence	0.39
PPV	0.59
NPV	0.71
LR+	2.23
LR-	0.62

Caucasian

Tabella 7.4: Risultati raccolti da ProPublica

Esaminiamo a questo punto i risultati raccolti per il gruppo "African-American". Analogamente a quanto fatto in precedenza confrontiamo i risultati contenuti nella Tabella 9.6 con quelli nella Tabella 9.7. Osserviamo che, anche questa volta, i valori riportati sono pressochè gli stessi con un False Positive Rate pari al 44.85% nel caso di ProPublica contro il 45% stimato da Aequitas. Nel caso del False Negative Rate, ProPublica riporta un valore pari al 27.99% in linea con il 28% di Aequitas.

Black defendants	Low	High	Percentage
Survived	990	805	0.49
Recidivated	532	1369	0.51
Total: 3696.00			

African-American

Tabella 7.5: Analisi ProPublica

Metrics	Values
False positive rate	44.85
False negative rate	27.99
Specificity	0.55
Sensitivity	0.72
Prevalence	0.51
PPV	0.63
NPV	0.65
LR+	1.61
LR-	0.51

African-American

Tabella 7.6: Risultati raccolti da ProPublica

Attribute Value	Group Size Ratio	False Discovery Rate	False Positive Rate	False Negative Rate
African-American	0.51	0.37	0.45	0.28
Asian	0	0.25	0.09	0.33
Caucasian	0.34	0.41	0.23	0.48
Hispanic	0.09	0.46	0.21	0.56
Native American	0	0.25	0.38	0.1
Other	0.05	0.46	0.15	0.68

Audit Results: Group Metrics Values

Tabella 7.7: Risultati raccolti da Aequitas

Infine, non è possibile fare ulteriori confronti tra le due analisi, in quanto nel "Bias Report" prodotto da Aequitas non vengono considerati i casi di recidiva violenta.

Conclusioni

Al termine di questo lavoro, possiamo concludere che è difficile stabilire con certezza chi abbia ragione. Questo dibattito ha visto contrapposti pareri molto discordanti tra loro. Da una parte coloro che difendono l'attendibilità dell'inchiesta condotta dai giornalisti di ProPublica, i quali sostengono che le analisi siano state fatte in modo completo e rigoroso. Dall'altra quelli che ritengono che le metriche utilizzate nello studio non siano significative e rappresentative nel contesto considerato. Gli stessi criticano anche le tecniche adottate da ProPublica ritenute poco accurate e approssimative a tratti. A questo proposito sono stati pubblicati numerosi articoli nella letteratura a favore dell'una o dell'altra parte, ma rimane ancora una questione aperta che, oltre a destare molto interesse nella gente, continua ad animare il dibattito tra esperti e scienziati.
Certo è che in numerosi stati degli USA vengono tuttora utilizzati strumenti per la valutazione del rischio, alcuni creati da società a scopo di lucro come la Northpointe e altri da organizzazioni non profit. Ci sono stati pochi studi indipendenti finora su queste valutazioni del rischio criminale. Tra questi rientra lo studio di ProPublica del 2016 di cui sono state riprodotte fedelmente le analisi all'interno di questa tesi ed ampliate con ulteriori analisi ricavate in parte con il toolkit Aequitas e in parte attraverso un'analisi intersezionale.
In conclusione, la maggior parte dei moderni strumenti per la valutazione del rischio sono stati originariamente progettati per fornire ai giudici uno strumento di supporto per aiutarli a valutare e a giudicare i casi che gli vengono sottoposti. Pertanto, tali strumenti non dovrebbero esercitare un potere decisionale sulla vita delle persone, altrimenti potrebbero diventare una reale minaccia per la libertà di ognuno di noi.

Bibliografia

[1] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016.

[2] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–7. [Online]. Available: https://doi.org/10.1145/3194770.3194776

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226. [Online]. Available: https://doi.org/10.1145/2090236.2090255

[4] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," 2018.

[5] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:13633339

[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021. [Online]. Available: https://doi.org/10.1177/0049124118782533

[7] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," 2016.

[8] G. Farnadi, B. Babaki, and L. Getoor, "Fairness in relational domains," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 108–114. [Online]. Available: https://doi.org/10.1145/3278721.3278733

[9] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," 01 2017.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, jul 2021. [Online]. Available: https://doi.org/10.1145/3457607

[11] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO '21. ACM, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1145/3465416.3483305

[12] C. R. Blyth, "On simpson's paradox and the sure-thing principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, 1972. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10482387

[13] R. Baeza-Yates, "Bias on the web," *Commun. ACM*, vol. 61, no. 6, p. 54–61, may 2018. [Online]. Available: https://doi.org/10.1145/3209581

[14] I. Northpointe, "Practitioner's guide to compas core," 2015.

[15] ProPublica, "Machine bias." [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[16] equivant, "Response to propublica: Demonstrating accuracy equity and predictive parity." [Online]. Available: https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

[17] ——, "Compas risk scales: Demonstrating accuracy equity and predictive parity." [Online]. Available: https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

[18] ProPublica, "How we analyzed the compas recidivism algorithm." [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[19] EIGE, "Intersezionalità." [Online]. Available: https://eige.europa.eu/publications-resources/thesaurus/terms/1050?language_content_entity=it#:~:text=L'analisi%20intersezionale%20mira%20a,e%20genere%20con%20altri%20motivi.

[20] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," 2019.

[21] Aequitas, "The bias report." [Online]. Available: http://aequitas.dssg.io/example.html#race-2

[22] GitHub, "Compas dataset." [Online]. Available: https://github.com/propublica/compas-analysis/