Enciclopedia dei dati digitali
Carlo Batini

Libro Terzo
L'etica dei dati digitali: l'equità
Versione 2
4 Luglio 2022

Indice

1. L'etica dei dati: introduzione all'equità p. 3
2. Modelli descrittivi, interpretativi, classificatori, predittivi, decisionali p. 14
3. Le tecniche di Machine learning p. 24
4. Come funziona una tecnica di Machine learning p. 28
5. L'equità nella filosofia, nelle scienze giuridiche e nella vita sociale p. 54
6. Quali principi etici chiediamo di rispettare a un modello basato sul Machine learning? p. 64
7. Cosa chiediamo ai modelli basati su tecniche di Machine learning? L'accuratezza p. 69
8. Cosa chiediamo ai modelli basati su Machine learning? L'equità, le equità P. 84
9. Questioni etiche che coinvolgono il modello, la tecnica, i dati e gli esseri umani p. 99
10. Metodi per mitigare la iniquità p.124

Appendice 1 - Ambienti didattici e ambienti di misurazione e mitigazione p.150
Appendice 2 - Definizioni dei termini più usati p.154
Appendice 3 – Per approfondire p.158

Questo è il terzo volume della Enciclopedia dei Dati Digitali.

Il primo volume "I dati sono una finestra sul mondo" è liberamente scaricabile dal link https://boa.unimib.it/handle/10281/301758

il secondo volume "I modelli dei dati ci aiutano a rappresentare e comprendere il mondo" è liberamente scaricabile dal link https://boa.unimib.it/handle/10281/301806

Capitolo 1. L'etica dei dati digitali: introduzione alla equità[1]

Nel 2015 Google Photos, uno dei siti di Google**,** ha creato una nuova funzione, che permetteva di classificare automaticamente le foto in categorie, sulla base del loro contenuto. Questa funzione è oramai (2022) utilizzabile anche su molti telefoni cellulari.

Il problema è che Google Photos, mentre assegnava correttamente per molte foto la classificazione, in alcuni casi (vedi Figura 1.1) classificava persone di etnia africana come *gorilla*.

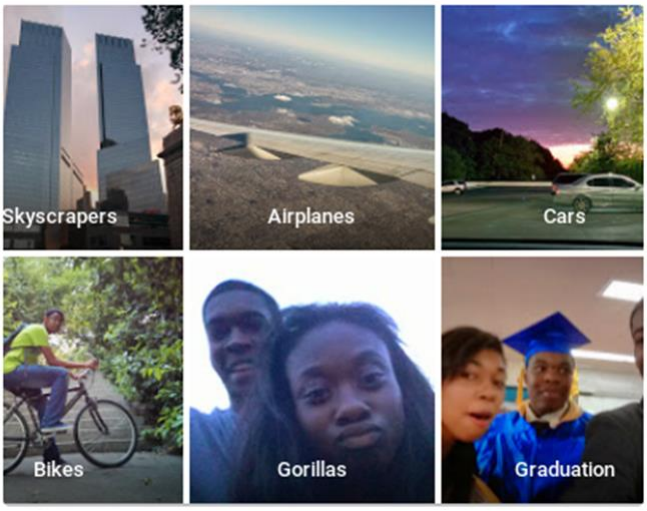


Figura 1.1 – Esseri umani o gorilla?

Come definiresti questo errore di classificazione?

Lo definirei razzismo…..

Sì, sono d'accordo con te. Google si giustificò affermando che l'errore non era dovuto al fatto che le immagini ritraessero persone di etnia africana, l'algoritmo non era prevenuto verso queste persone; l'errore era dovuto al fatto che l'immagine fosse scarsamente luminosa. Insomma, sarebbe potuto accadere anche per altre etnie.

Direi che non è una buona scusa: direi, come i veneti, "pezo el tacon del buso"

Certamente! L'errore che commetteva l'algoritmo derivava solo indirettamente dal colore della pelle, ma era appunto un errore che andava corretto prima.

[1] In questo volume, come nei precedenti, io fingo di interagire con un interlocutore, che ogni tanto prende la parola, fa delle osservazioni o delle domande; gli interventi del mio interlocutore sono evidenziati in *corsivo*. Il lettore tenga presente che sono evidenziate in corsivo nel testo anche *singole parole o insiemi di parole considerate rilevanti.* Infine, quando uso il voi, mi rivolgo all'insieme dei mei lettori.

La Figura 1.2 mostra due traduzioni effettuate da Google Translator di frasi dalla lingua inglese alla lingua italiana.

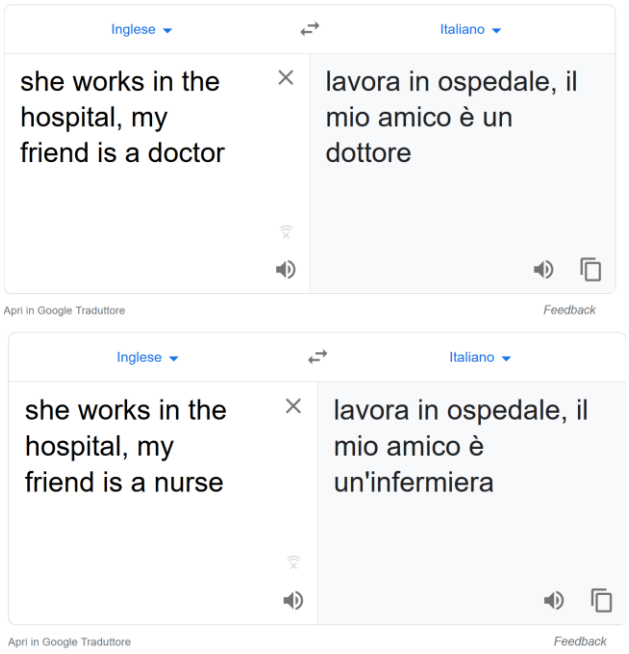


Figura 1.2 – L'equità di genere

Google Translator adotta una tecnica di apprendimento automatico (nel libro useremo prevalentemente il termine inglese *Machine learning,* con significato equivalente*)*, che apprende le traduzioni attraverso la elaborazione di un grandissimo numero di frasi che sono espresse in diverse lingue.

Ma dove trova queste traduzioni già fatte?

Le trova ad esempio in Wikipedia, in cui lo stesso concetto è descritto spesso in diverse lingue; esistono poi istituzioni come la Unione Europea che sono obbligate a pubblicare i documenti ufficiali in tutte le lingue dei Paesi aderenti.

Ebbene, cosa noti nelle traduzioni in Figura 1.2?

Sono strane queste traduzioni, termini con genere maschile, ad esempio "amico", vengono coniugati con termini sia maschili che femminili (dottore e infermiera). L'errore evidente sta nell'assumere che i dottori siano maschi e le infermiere siano femmine.

Bene, e come definiresti questo errore?

Direi che Google Translator ….fa un errore di genere; e poi, che è prevenuto verso le donne…

Certo, è così, ma entrando un po' più nel merito, anche in questo caso si può dire che la tecnica che apprende dalle traduzioni *non fa che fotografare una caratteristica delle lingue*, in cui storicamente, accanto al termine maschile, ad esempio "dottore", fa fatica ad affermarsi un termine di genere femminile, ad esempio dottoressa.

Non è sempre vero quello che dici, ad esempio sono oramai diffusamente utilizzati termini come senatrice, rettrice, ministra…

Sì, questo è vero, ma Google Translator impara dalle traduzioni che trova, impara dal passato, in forma acritica, e quei termini appaiono solo in testi molto recenti, e quindi molto rari nei documenti sulla base dei quali apprende a tradurre.

In Figura 1.3 vediamo il caso dello strumento Compas utilizzato negli Stati Uniti per aiutare i giudici a decidere se concedere o meno la libertà provvisoria ai detenuti in attesa di giudizio[2].

Compas basa la sua decisione sui dati storici disponibili riguardanti il comportamento dei detenuti nel passato: per ogni detenuto cui è stata concessa la libertà provvisoria, è noto se abbia commesso recidiva o meno; sulla base di questi dati, Compas determina il grado di rischio che possano commettere recidiva i detenuti con caratteristiche simili ai detenuti del passato, che chiedono *oggi* la libertà provvisoria.

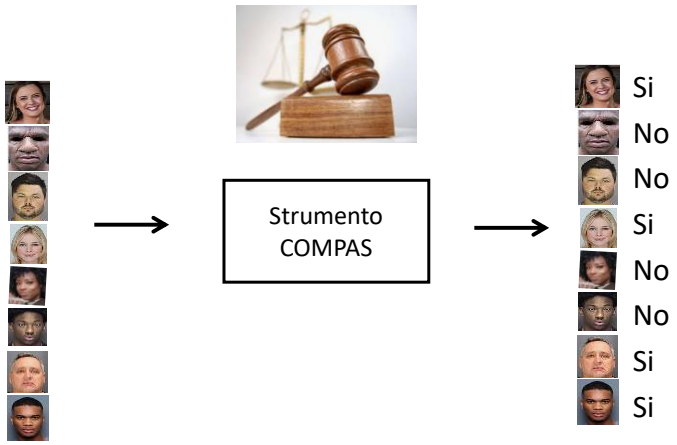


Figura 1.3 - Lo strumento Compas utilizzato negli Stati Uniti in diverse contee per assistere i giudici nella concessione della libertà provvisoria

Ebbene, il giornale d'inchiesta ProPublica (www.Propublica,org, verificato nel gennaio 2022) produsse nel 2016 una analisi in cui mostrava come Compas tendesse a sopravvalutare il rischio, e *quindi a concedere in misura minore la libertà provvisoria ad alcune categorie di detenuti, tra cui gli Afro-Americani*. In questo caso, come definiresti Compas alla luce della denuncia di ProPublica?

Direi che Compas…. è discriminatorio, non equo, verso gli Afro-Americani.

Certo. In tutte le società esistono diseguaglianze, di reddito, nei diritti politici, nell'accesso ai servizi sociali, e in tanti altri campi della vita civile, disuguaglianze che nel seguito chiameremo con il termine *discriminazioni*. Da tempo immemorabile gli Stati e le Istituzioni internazionali definiscono nelle loro leggi costitutive principi di eguaglianza e non discriminazione; nella Figura 1.4 ho riprodotto i primi due articoli della Dichiarazione universale dei diritti umani

[2] Attenzione, questo esempio verrà utilizzato molto spesso nel libro….

(nota nella figura che la parola “umani” veniva tradotta fino a pochi anni fa, e viene tuttora spesso tradotta, con il termine “dell’uomo”..).

Dichiarazione universale dei diritti dell'uomo
(Approvata dall'assemblea delle Nazioni Unite il 10 dicembre del 1948)
Articolo 1
Tutti gli esseri umani nascono liberi e uguali in dignità e diritti. Sono dotati di ragione e di coscienza e devono agire in uno spirito di fraternità vicendevole.
Articolo 2
Ognuno può valersi di tutti i diritti e di tutte le libertà proclamate nella presente dichiarazione, senza alcuna distinzione di razza, di colore, di sesso, di lingua, di religione, d'opinione politica e di qualsiasi altra opinione, d'origine nazionale o sociale, che derivi da fortuna, nascita o da qualsiasi altra situazione. Inoltre non si farà alcuna distinzione basata sullo statuto politico, amministrativo o internazionale del paese o del territorio a cui una persona appartiene, sia detto territorio indipendente, sotto tutela o non autonomo, o subisca qualunque altra limitazione di sovranità.

Figura 1.4 - La dichiarazione universale dei diritti umani

In Figura 1.5 è riportato l’Articolo 3 della Costituzione Italiana, in cui viene affermato il principio di eguaglianza di tutti i cittadini, senza distinzione di sesso, di razza, di lingua, di religione, di opinioni politiche, di condizioni personali e sociali. Anche questo principio va nella direzione di stabilire condizioni di equità tra tutti i cittadini.

Art. **3**. Tutti i cittadini hanno pari dignità sociale e sono eguali davanti alla legge, senza distinzione di sesso, di razza, di lingua, di religione, di opinioni politiche, di condizioni personali e sociali.

Figura 1.5 – La Costituzione italiana

Nota come in questo caso venga adottato un termine diverso e più specifico di *uomo* o *persona,* per indicare le persone che hanno diritti, il termine *cittadino*. E nota inoltre che rispetto alla dichiarazione universale dei diritti umani dell’ONU, qui vengono espressamente individuate *le caratteristiche* delle persone che storicamente hanno causato le discriminazioni, e cioè il sesso (che oggi chiamiamo *genere*), la razza (che oggi chiamiamo *etnia*), e tutte le altre menzionate nell’articolo.

In questo libro vogliamo investigare l’equità come categoria dell’etica; non la vogliamo indagare in generale, ma con riferimento ai dati digitali e alle tecniche di Machine learning sviluppate nell’ambito della Intelligenza Artificiale. L’Intelligenza Artificiale è una disciplina molto più ampia del Machine learning, e comprende la logica simbolica, la robotica,

l'elaborazione del linguaggio naturale, i modelli conessionistici, le reti neurali, le reti di agenti intelligenti.

Iniziamo ora la nostra indagine sull'etica, dando uno sguardo generale alla relazione tra etica e dati digitali; subito dopo, parleremo di etica e tecniche di Machine learning.

1.1 Etica e dati digitali

L'abbiamo visto nel primo libro della Enciclopedia, i dati digitali sono un artefatto che è al tempo stesso tecnologia e rappresentazione del mondo. Le tecnologie dei telefoni mobili, dell'Internet delle cose e delle reti sociali nascono e si diffondono con l'obiettivo di rappresentare potenzialmente ogni aspetto del mondo. In tal modo, assistiamo ad una progressiva commistione tra i due mondi dell'analogico e del digitale, che porta a rendere più complessa la definizione delle responsabilità etiche e del libero arbitrio della persona e l'influenza che su di esse esercitano le tecnologie del Machine learning.

In tale contesto di pervasività dei dati digitali nella vita delle comunità e degli individui, i ricercatori di diverse discipline hanno iniziato a interrogarsi sulla relazione esistente tra i dati digitali e l'etica. Per Wikipedia, l'etica è una branca della filosofia che studia i fondamenti razionali che permettono di assegnare ai comportamenti umani uno status deontologico, ovvero distinguerli in buoni, giusti, leciti, rispetto ai comportamenti ritenuti ingiusti, illeciti, sconvenienti o cattivi secondo un ideale modello comportamentale, ad esempio una data morale.

Non mi arrischio a trattare il tema dell'etica in termini generali, non ho gli strumenti. Piuttosto, propongo una lista di *determinanti*, cioè temi e aspetti della nostra vita che, a mio parere, influiscono su un comportamento etico quando vengono usati i dati digitali:

1. *Data divide*, è l'ineguaglianza di natura economica, sociale o culturale che porta a ineguaglianze nelle classi sociali riguardo all' accesso ai dati, all'uso dei dati, al valore che i dati hanno per l'utente.
2. *Trasparenza (dei dati digitali*) è la capacità dei dati digitali di descrivere in modo comprensibile a tutti, indipendentemente dalla loro cultura digitale e non, il fenomeno (ad esempio, un evento, una organizzazione, una legge) che essi rappresentano.
3. *Accessibilità*, esprime la possibilità data agli utenti di poter fruire dei dati digitali di loro interesse, sia in termini di tecnologie di accesso (ad es. la rete Internet) che in termini di comprensione del loro significato.
4. *Accountability*, o *assunzione di responsabilità*, esprime la esistenza e la messa a disposizione di dati e informazioni per identificare chi, a partire dai dati disponibili, abbia preso una decisione o abbia effettuato una azione cui è connessa una responsabilità, e le ragioni di tale decisione. E' anche la *verificabilità*, la capacità di esaminare e valutare in profondità comportamenti o azioni riferiti a dati digitali.
5. *Generalizzazione verso Personalizzazione*, individuazione di un equilibrio tra la messa a disposizione di dati generali, validi per tutti, ovvero la personalizzazione e l'adattamento dei dati verso specifiche comunità o individui.

6. *Qualità,* proprietà dei dati di essere corretti, completi, aggiornati, essendo in tal modo aderenti alla realtà. Affronterò questo aspetto nel Libro IV della Enciclopedia.
7. *Privacy*, la caratteristica dei dati personali (ad es. codice fiscale, cognome, nome) e sensibili (ad es. opinioni politiche, religione) di essere accessibili solo a soggetti autorizzati (ad es. pubbliche amministrazioni) e per particolari usi.
8. *Condivisione o apertura*, che dà la possibilità a tutti di accedere ai dati digitali e a non trattarli come un bene privato della persona o organizzazione che li produce.

Accanto a questi determinanti ve ne sono altri, che discuteremo alla fine della prossima Sezione 1.3, in cui introduciamo i concetti principali del libro: i modelli classificatori, predittivi e decisionali, e le tecniche di apprendimento automatico che creano tali modelli.

1.3 Modelli, tecniche, determinanti etici

Torniamo al caso dello strumento Compas, che aiuta i giudici a decidere se concedere o meno la libertà provvisoria a detenuti che la chiedano, vedi Figura 1.6.

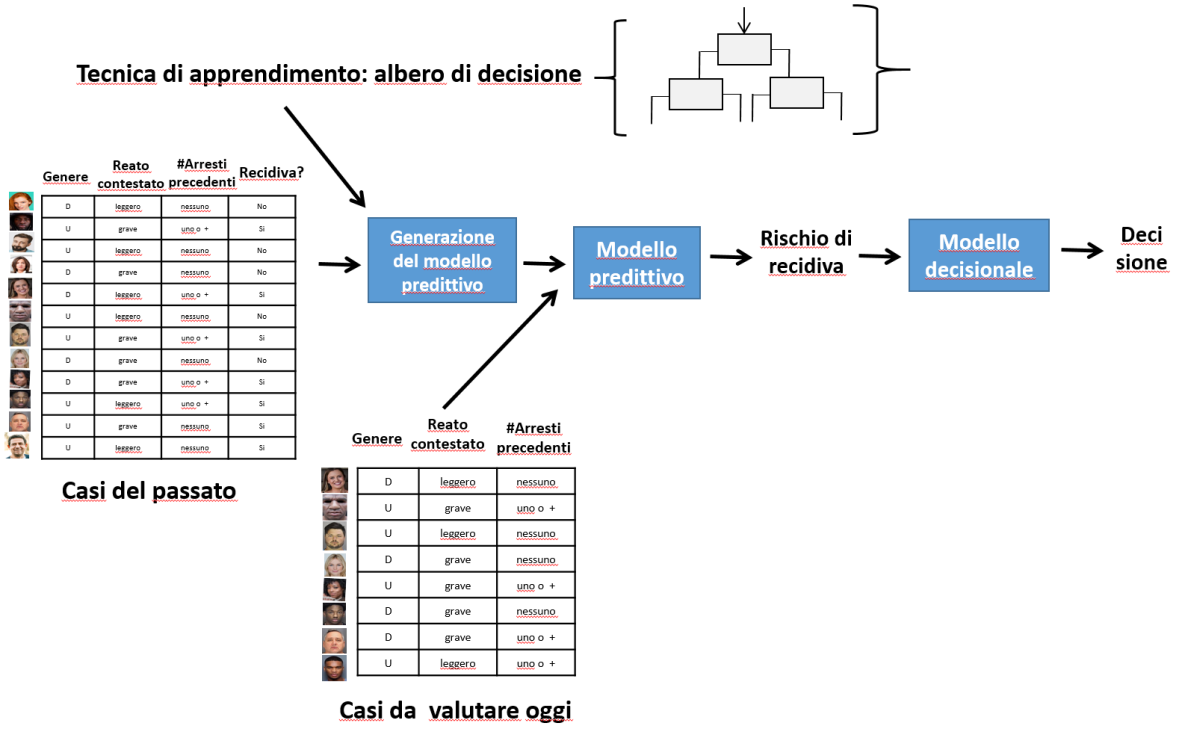


Figura 1.6 – Modelli predittivi e tecniche di Machine learning
per decidere se concedere o meno la libertà provvisoria

La decisione viene presa cercando di prevedere quale sarà il rischio che i detenuti una volta liberati possano commettere recidiva. Il rischio è determinato mediante un *modello predittivo*, che, in sostanza, partendo da un insieme di dati in ingresso riferiti ai detenuti che oggi chiedono la libertà provvisoria, formula una previsione sul rischio che tali detenuti possano commettere recidiva, possano commettere, cioè, nuovi reati.

Come può un modello predittivo formulare una tale previsione? Ha la sfera di cristallo?

Il rischio viene calcolato sulla base del comportamento tenuto dai detenuti *a cui è stata concessa nel passato la libertà provvisoria*. Di ogni detenuto si assume di conoscere le seguenti ca:

- il genere (uomo o donna),
- il tipo di reato contestato (leggero o grave),
- il numero di arresti precedenti al momento in cui il detenuto è stato liberato (che assumeremo possa avere i due valori: zero, uno o +),

e, infine, se abbia commesso o meno recidiva successivamente alla liberazione.

Ciò viene fatto mediante una *tecnica di apprendimento*, che, appunto, impara dal passato. Ad esempio, un albero di decisione, vedi Figura 1.7, è una tecnica di apprendimento che, attraverso un insieme di domande (ad es. è uomo o donna?), permette di distinguere per ogni detenuto del passato a quale gruppo appartenga.

Per ogni gruppo di detenuti con gli stessi valori delle tre caratteristiche, ad es: [uomo, reato grave, nessun arresto precedente], si calcola quanti abbiano commesso recidiva (nel seguito n1) e quanti no (n2) , e si assegna al gruppo *rischio alto o basso* a seconda che n1 sia maggiore di n2, o meno, vedi Figura 1.7.

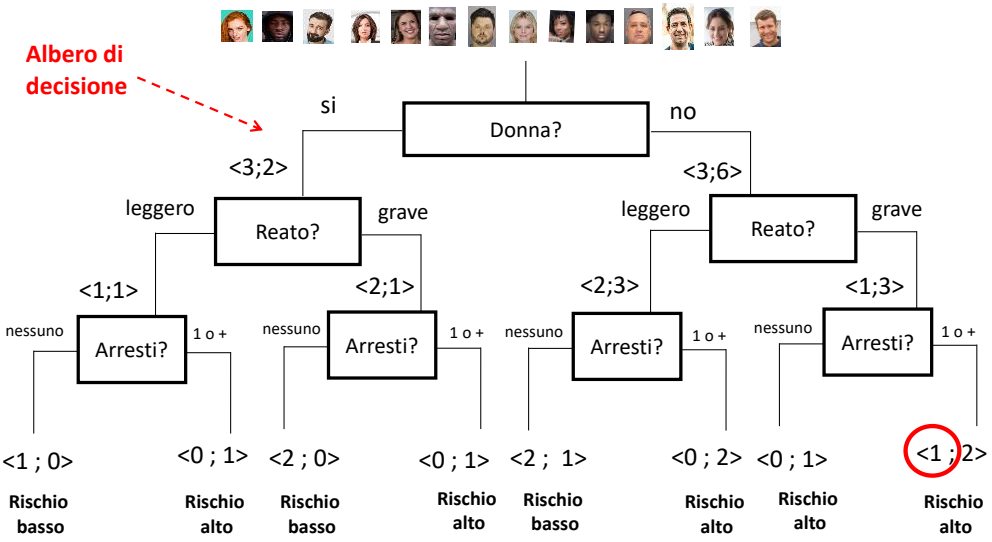


Figura 1.7 – Un albero di decisione costruito a partire dalle caratteristiche di quattrodici detenuti del passato

Fai attenzione, nella Figura 1.7, al fatto che, ad esempio, tra gli uomini che hanno commesso reati gravi e che hanno subito 1 o + arresti, ce ne è uno, cerchiato in rosso che non ha commesso recidiva, e a cui viene assegnato rischio alto. Nel libro chiameremo questi casi *falsi positivi* (cioè elementi che hanno rischio basso ma a cui è assegnato rischio alto).

Una volta costruito il modello predittivo, per prendere la decisione sui detenuti che oggi chiedono la libertà provvisoria, si determina il rischio di recidiva:

1. valutando a quale gruppo appartenga il detenuto, e
2. associando al detenuto il rischio corrispondente al gruppo.

A questo punto il giudice decide sulla base del rischio associato al detenuto, in questo caso la decisione più naturale è rischio basso → libertà si, rischio alto → libertà no.

Un momento: ma veramente la decisione viene presa in questo modo? Veramente per un evento così importante per un individuo, la decisione viene presa con un ragionamento così….superficiale?

Sì, comprendo il tuo sconcerto, ma questo è il modo in cui funzionano i modelli predittivi basati su tecniche di apprendimento. Comunque, avremo modo di ragionare su questa tua importante domanda nel resto del libro, soprattutto nell'ultimo capitolo.

Adesso ho una domanda per te. Guarda nuovamente la Figura 1.7. Osserva tutti i casi: c'è qualche elemento falso negativo, cioè che è ad alto rischio, e che però è classificato a basso rischio, e perciò viene liberato?

La risposta nella prossima pagina.

Risposta – Certo che c'è un falso negativo, è quello cerchiato di rosso in Figura 1.8. Infatti appartiene a un gruppo che ha due elementi a basso rischio, per cui al gruppo è assegnato basso rischio, pur avendo l'elemento alto rischio.

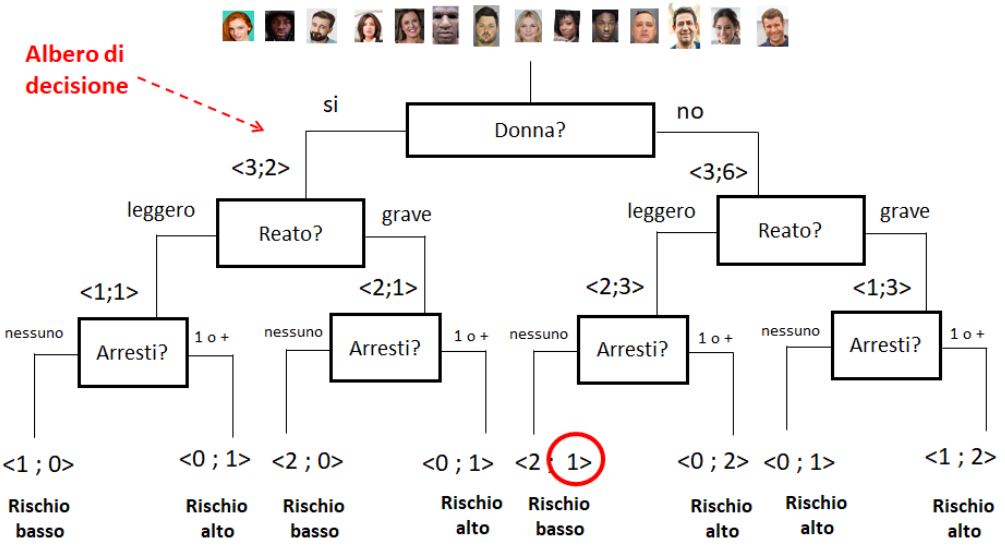


Figura 1.8 – L'elemento falso negativo è quello cerchiato di rosso

Comune a tutti i modelli predittivi che studiamo in questo libro è il fatto che vengano generati a partire da una tecnica di apprendimento che impara da casi passati. Modello predittivo e tecnica di apprendimento sono concetti centrali il questo libro.

E centrali nel libro sono i tre ulteriori determinanti dell'etica che possiamo associare ai modelli predittivi: l'accuratezza, l'equità, e la spiegabilità (explicability in inglese).

9. *L' accuratezza* è la capacità di modello (e della tecnica di apprendimento che lo ha generato) di minimizzare gli errori per gli scopi per cui è utilizzato; per esempio, nel caso Compas l'accuratezza corrisponde al fatto che tra i detenuti non ci siano falsi positivi, cioè detenuti che nel passato non hanno commesso recidiva eppure sono stati classificati ad alto rischio, perché hanno prevalso quelli che hanno commesso recidiva, né falsi negativi, il caso opposto.
10. *L'equità* (fairness in inglese, anche imparzialità, obiettività, oggettività, assenza di distorsioni o bias, dall'inglese), la proprietà del modello predittivo di non fare discriminazioni nelle predizioni tra diversi gruppi sociali (ad esempio, uomini o donne) o diversi individui.
11. *La spiegabilità* è la capacità del modello e della tecnica di far comprendere il processo di calcolo che ha portato all'apprendimento. Ad esempio gli alberi di decisione hanno alta speigabilità, perché è facile seguire i percorsi di decisione nell'albero, dalla radice alle foglie.

Nel seguito saremo soprattutto interessati a indagare la accuratezza e la equità.

1.4. Organizzazione del libro

In questo libro ci occupiamo della accuratezza ed equità dei modelli predittivi per dati digitali, con particolare riferimento ai modelli che utilizzano tecniche di apprendimento automatico (*Machine learning* in inglese).

La produzione dei modelli basati su apprendimento ha ricevuto impulso negli ultimi anni dalla disponibilità di grandi quantità di dati che descrivono fenomeni di interesse, e dallo sviluppo di tecniche nell'ambito della Intelligenza Artificiale e del Machine Learning.

Con il termine *Intelligenza Artificiale* intendo, in accordo con Wikipedia Inglese, l'intelligenza mostrata dalle macchine, come opposta alla intelligenza naturale mostrata dagli animali, inclusi gli esseri umani. I principali testi sulla Intelligenza Artificiale la definiscono come un qualunque sistema che percepisce l'ambiente in cui opera, ed effettua azioni che massimizzano la possibilità di raggiungere i propri obiettivi.

Il Machine learning è quella parte della Intelligenza artificiale che genera modelli, nel senso introdotto poco fa, mediante l'uso di tecniche basate su apprendimento.

In particolare le applicazioni del Machine learning di maggiore rilevanza sono (senza pensare di averle individuate tutte!):

1. Il riconoscimento e classificazione di immagini
2. La traduzione di testi tra lingue (es. Google Translator)
3. L'applicazione della legge (Law enforcement) e la previsione di attività criminali potenziali (predictive policing)
4. La Giustizia Penale e Civile: ad es. le sentenze penali e civili, la concessione della libertà provvisoria e altre decisioni.
5. L'auto a guida autonoma
6. I sistemi di raccomandazione che usiamo quando ci viene suggerito un acquisto da siti come Amazon o eBay.
7. La pubblicità on line
8. I giochi di Scacchi e di Go
9. Il riconoscimento facciale
10. La traduzione di testi e Chatboat (interfacce vocali)
11. Gli investimenti finanziari
12. I motori di ricerca
13. I filtri anti spam
14. I sistemi militari e di puntamento
15. I sistemi di sorveglianza.

I Capitoli da 2 a 4 approfondiscono i modelli e le tecniche di Machine learning.

Nel Capitolo 2 viene definito il concetto di modello e vengono indagati i diversi tipi di modelli trattati nelle Scienze, in particolare i modelli classificatori e predittivi, per la cui costruzione possiamo utilizzare tecniche basate su Machine learning. Il Capitolo 3 descrive le diverse

tipologie di tecniche di Machine learning. Il Capitolo 4 si focalizza sul Machine learning a partire da esempi, e ne spiega il funzionamento.

I capitoli 5 e 6 discutono dell'equità in generale e delle categorie etiche nella Intelligenza Artificiale.

Il Capitolo 5 discute come l'equità sia trattata nella filosofia e nelle scienze giuridiche. Nel Capitolo 6 indaghiamo alcune delle iniziative che più hanno influenzato la individuazione di categorie etiche rilevanti nella Intelligenza Artificiale.

I capitoli 7 e 8 approfondiscono la accuratezza e equità nel Machine learning.

A partire dal Capitolo 7 iniziamo a indagare le proprietà dei modelli che usano tecniche di Machine learning, nel Capitolo 7 parleremo di accuratezza e nel Capitolo 8 di equità. Nel Capitolo 7 vengono mostrate diverse tipologie di accuratezza, come, ad esempio, la sensibilità e la specificità, spesso citate nei test che rivelano patologie o malattie da virus. Il Capitolo 8 introduce diverse forme di equità nel Machine learning.

I capitoli 9 e 10, i più importanti del libro, approfondiscono i temi etici nei modelli predittivi e decisionali considerati insieme, e le regole che possiamo seguire per mitigare le iniquità, sia nei modelli e tecniche, sia nella società.

Nel Capitolo 9 vedremo come l'equità dei modelli sia influenzata da diversi fattori, quali i dati di ingresso, la tecnica utilizzata, la presenza di feedback o retroazioni nell'utilizzo del modello nel tempo. Nel Capitolo 10, infine, vedremo come sia possibiie mitigare la iniquità nel Machine learning, e scopriremo due tipi di regole di mitigazione, le regole tecnologiche, che fanno riferimento al modello, i dati digitali in ingresso e la tecnica di apprendimento, e le regole sociali, che riguardano tutti noi he apparteniamo alla società.

L'Appendice 1 fornisce i link ai principali siti che forniscono moduli didattici sulla accuratezza e la equità, e ambienti programmativi per la loro misurazione e mitigazione. L'Appendice 2 fornisce le definizioni dei termini più usati nel libro, e l'Appendice 3 indica alcuni testi per approfondimenti.

I precedenti argomenti vengono affrontati partendo da diversi studi di caso ed esempi, tra essi il caso dominante è il caso Compas sulla concessione della libertà provvisoria. Il testo è corredato anche da diversi esercizi; lo sforzo è fare in modo che il lettore, per esempio uno studente di scuola secondaria, possa diventare protagonista attivo nell'apprendere i concetti esposti nel libro.

Buona lettura!

Capitolo 2 - Modelli descrittivi, classificatori, predittivi, decisionali

La grande disponibilità di dati digitali, di cui abbiamo parlato a lungo nel Libro 1 della Enciclopedia [3] rende possibile elaborare una miriade di differenti descrizioni ed elaborazioni della realtà rappresentata dai dati digitali, che chiamiamo con il nome unificante di *modelli*.

Wikipedia fornisce come definizione di modello la seguente: un modello è, in Fisica ma anche in altri settori della conoscenza, una rappresentazione concettuale del mondo reale o di una sua parte, capace di *spiegare* un determinato fenomeno. In Fisica, o comunque nella Scienza in genere, si parla di modelli fisici che descrivono i fenomeni reali.

Ad esempio le equazioni di Maxwell, rappresentate in Figura 2.1, esprimono l'evoluzione temporale e i vincoli a cui è soggetto un campo elettromagnetico in relazione alle distribuzioni di carica e corrente elettrica da cui è generato.

Legge di Maxwell	**Significato**	**Equazione differenziale**
Legge di Gauss	il flusso del campo elettrico attraverso una superficie chiusa è proporzionale alla carica interna alla superficie	$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$
Legge di Gauss per il magnetismo	Non ci sono cariche magnetiche analo-ghe alle cariche elettriche. Il flusso del campo magnetico attraverso una superficie chiusa e' pari a zero.	$\nabla \cdot \mathbf{B} = 0$
Equazione di Maxwell–Faraday	il lavoro per unità di carica necessario a spostare una carica intorno a una spira chiusa è pari al tasso di diminuzione del flusso magnetico attraverso la superficie racchiusa	$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
Legge di Ampere	il campo magnetico indotto intorno a un circuito chiuso è proprorzionale alla corrente elettrica più la corrente di spostamento (proporzionale al tasso di cambiamento del flusso) attraverso la superficie chiusa	$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)$

Figura 2.1 – Un modello della Fisica: Le equazioni di Maxwell

La grande potenza di un modello come le leggi di Maxwell sta nel fatto che sono di applicazione generale: qualunque fenomeno elettromagnetico rispetta le quatto leggi di Maxwell, per cui, se ad esempio vogliamo progettare e realizzare una antenna, oppure una rete di trasmissione, possiamo utilizzare le quattro leggi di Maxwell senza ogni volta porci da capo il problema di come procedere.

Un *modello matematico* è, secondo Wikipedia, una rappresentazione quantitativa di un fenomeno naturale. Ad esempio, se vogliamo rappresentare la relazione che esiste tra le lunghezze dei lati di un qualunque triangolo rettangolo (Vedi Figura 2.2), il teorema di

[3] Il libro è liberamente scaricabile al link https://boa.unimib.it/handle/10281/301758

Pitagora (vedi Figura 2.3) ci dice che la somma dei quadrati delle lunghezze dei due lati che formano l'angolo di 90 gradi, è pari al quadrato della lunghezza della ipotenusa.

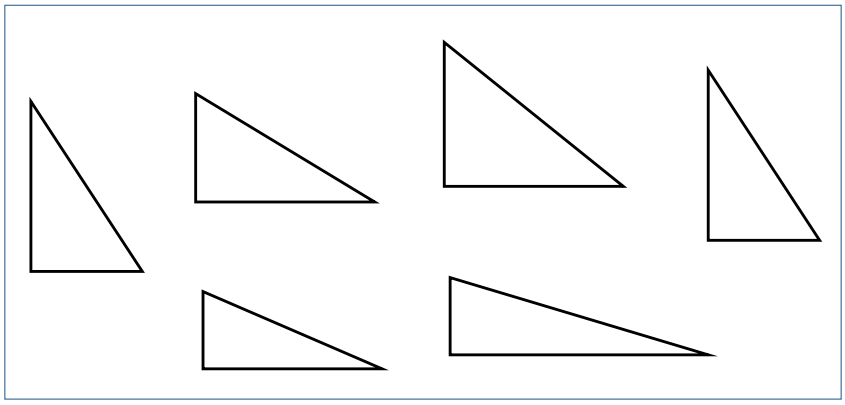


Figura 2.2 - Un insieme di triangoli rettangoli

$$\text{Lunghezza del primo lato}^2 + \text{Lunghezza del secondo lato}^2 = \text{Lunghezza della ipotenusa}^2$$

Figura 2.3 – Il teorema di Pitagora

Anche i modelli matematici esprimono proprietà generali, applicabili in un insieme illimitato di casi; se, ad esempio, noi conosciamo di un qualunque triangolo rettangolo le lunghezze di uno dei due lati e della ipotenusa, possiamo calcolare la lunghezza del secondo lato.

I modelli più utilizzati nell'informatica, nella statistica e nella intelligenza artificiale sono i modelli descrittivi, modelli interpretativi, modelli classificatori, modelli predittivi e i modelli prescrittivi o decisionali.

2.1 Modelli descrittivi

I modelli descrittivi hanno lo scopo, come dice il termine, di descrivere un fenomeno del mondo reale; non ambiscono a descrivere leggi applicabili in modo generale nella realtà, piuttosto ambiscono a esprimere le proprietà rilevanti di un fenomeno della realtà.

Consideriamo ad esempio le persone che vivono in Italia; noi siamo interessati a sapere quante sono (attraverso operazioni di conteggio sui risultati di un censimento), come sono suddivise nelle varie regioni (ad esempio attraverso operazioni di percentuale sul totale della popolazione), quale sia l'età media, quale sia l'età mediana, come siano suddivise per fascia di età e regione di residenza (mediante operazioni di aggregazione). La statistica è la disciplina che fornisce gli strumenti concettuali per costruire modelli descrittivi. Si usa dire che i modelli descrittivi operano sul *cosa è*. In Figura 2.4, che riprende una figura del Libro 1 della Enciclopedia, mostriamo alcune statistiche descrittive su persone decedute, casi positivi e altri indicatori per le regioni italiane.

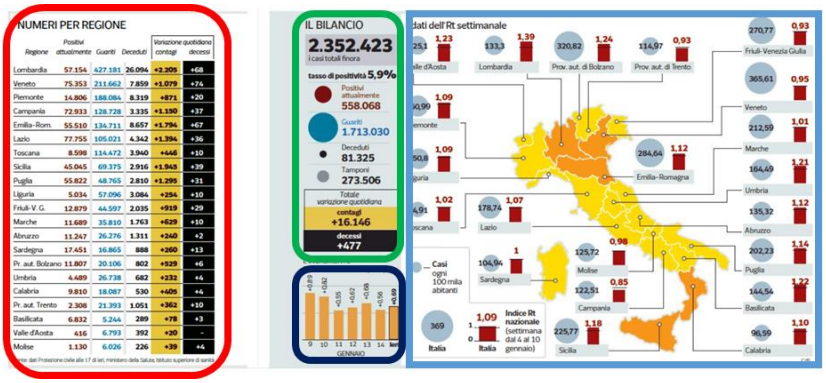


Figura 2.4 – Statistiche descrittive nella epidemia dovuta al virus Covid

Non svilupperemo nel seguito del libro i modelli descrittivi.

2.2 Modelli interpretativi o diagnostici

I modelli interpretativi indagano su un fenomeno, individuando legami statistici o legami causali tra i dati (ad esempio la relazione che sussiste tra titolo formativo conseguito nel percorso scolastico e tipo di occupazione), al fine di interpretare il fenomeno e ricostruirne le cause.

Ad esempio, in Figura 2.5 vediamo una persona che ha mal di testa. Perché certe volte ci viene mal di testa? Perché siamo stanchi, perché abbiamo mangiato male, perché siamo in un posto molto caldo? I modelli interpretativi forniscono risposte sulle cause degli eventi e degli stati del mondo. Si usa dire che i modelli interpretativi descrivono il *perchè è*.

Figura 2.5 – Perché ho mal di testa?

Anche i modelli interpretativi non saranno tratti nel resto del libro.

2.3 Modelli classificatori o classificazioni

Alcune applicazioni per telefoni smart permettono di fotografare un albero, classificandolo, partendo dalla forma che ha la chioma, in: abete, acero, larice, quercia, e così via.

Questo è un tipico problema di classificazione. I *modelli classificatori o classificazioni* sono modelli che hanno lo scopo di stimare il valore di una caratteristica y (ad esempio il tipo di albero) a partire dai valori che assumono un insieme di caratteristiche x1, x2, .., xn (ad esempio l'altezza dell'albero, il colore del tronco, la forma delle foglie, l'ampiezza della chioma, ecc.).

Semplificando, supponiamo di avere a disposizione un insieme di foto di soli larici e querce (vedi Figura 2.5), per le quali sappiamo già se rappresentino un larice o una quercia; un modello classificatorio ha la capacità, partendo da nuove foto di larici o querce, di distinguere se sia un larice o una quercia.

Ma come fa un modello classificatorio a … capirlo?

Possiamo creare il modello classificatorio utilizzando una tecnica di apprendimento che apprende se un albero sia larice o quercia partendo dagli esempi disponibili, i primi tre di Figura 2.6.

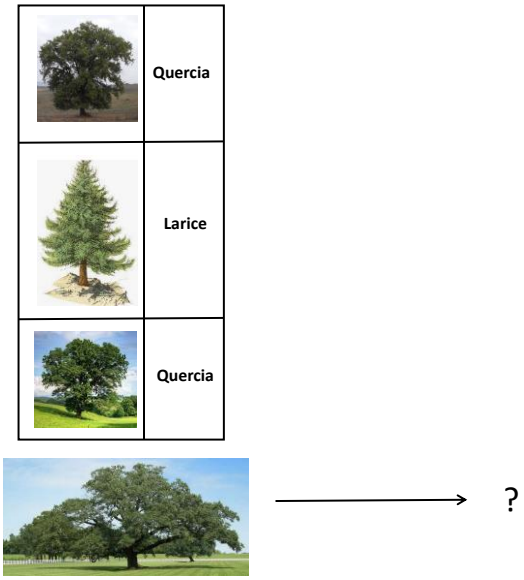


Figura 2.6 – Classificazione in larici e querce

L' esempio di Figura 2.6 ci dice che non è semplice individuare quali sono le caratteristiche da considerare, per decidere nel caso del nuovo albero che ha la chioma così ampia; se alla tecnica di apprendimento non forniamo esempi di alberi con chioma larga, per la tecnica e per il modello sarà come tirare una moneta …

Le classificazioni creano un ordine (essere larice o quercia, essere mela o pera o susina, ecc.) a partire da un insieme disordinato (un bosco di larici e querce insieme, una scodella con mele e pere e susine).

2.4 Modelli predittivi

I modelli predittivi sono modelli classificatori in cui la caratteristica y risultato della classificazione riguarda *un fenomeno che accadrà nel futuro*; è per tale ragione che li introduciamo con un termine distinto rispetto ai modelli classificatori. Guardiamo la Figura 2.7, in cui ho rappresentato una eclissi di sole e una eclissi di luna.

Il problema di prevedere le eclissi esista da millenni, se lo posero per primi gli astronomi babilonesi, che durante l'VIII ed il VII secolo a.C. svilupparono un sistema empirico di previsione delle eclissi di sole e di luna. Questo sistema si basava sulle osservazioni periodiche delle eclissi nel tempo; interpolando le date delle eclissi passate, riuscirono a prevedere con buona approssimazione le eclissi future. Possiamo dire che gli astronomi babilonesi costruirono un modello predittivo basato sul numero di anni che passava regolarmente tra una eclissi e l'altra.

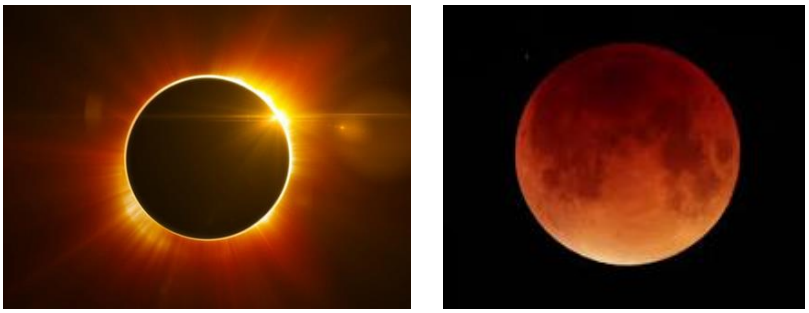


Figura 2.7 – Predire quando si verificheranno le eclissi di sole

Affrontiamo ora un problema che solo recentemente è stato risolto con un modello predittivo. La Figura 2.8 mostra il risultato di una ricerca che noi effettuiamo spesso quando vogliamo comprare un biglietto aereo: *fissato il giorno di partenza e fissati gli aeroporti di partenza e arrivo*, vogliamo trovare il volo più economico dal primo al secondo aeroporto, *assumendo che acquisti oggi il biglietto*. Questo problema è facilmente risolvibile avendo a disposizione tutti gli orari di partenza e arrivo delle compagnie aeree, le tratte percorse e i costi dei biglietti.

Basta infatti avere la pazienza di collegare tutte le tabelle dei voli e dei costi dei biglietti coinvolte dal nostro problema, mediante operazioni analoghe a quelle viste nel Volume 2 della Enciclopedia dedicato ai modelli dei dati; per esempio, per i voli da Milano Linate a Los Angeles, dobbiamo prendere in considerazione sia i voli diretti, sia i voli che permettono di arrivare a Los Angeles con più scali. Otteniamo in questo modo una nuova tabella con l'elenco dei voli diretti, i voli con uno scalo intermedio, quelli con due scali intermedi, ecc. nel giorno di nostro interesse, e relativi prezzi; a questo punto ordiniamo i voi dal volo con il prezzo più basso al volo con il prezzo più alto e otteniamo le informazioni rappresentate in Figura 2.8.

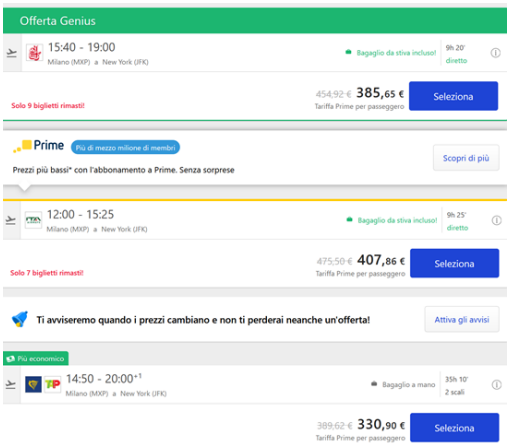


Figura 2.8 – Fissato il giorno e gli aeroporti di partenza e arrivo, trovare il volo più economico.

Ebbene, noi potremmo avere una *esigenza diversa*: fissato il giorno di partenza e gli aeroporti di partenza e arrivo, vogliamo *prevedere il giorno, tra oggi e il giorno della partenza, in cui conviene acquistare il biglietto.*

Se ci pensate, ci troviamo davanti a una sfera di cristallo: come facciamo a prevedere il futuro, come facciamo a prevedere come evolverà il prezzo dei biglietti da ora al giorno della partenza? Possiamo immaginare alcune leggi generali sul prezzo del biglietto: man mano che si avvicina il giorno il prezzo tenderà ad aumentare, e poi, nella imminenza della partenza, probabilmente i prezzi crolleranno; ma queste sono leggi generali, noi abbiamo bisogno di una previsione molto più fine su come le compagnie aeree fanno variare i prezzi dei biglietti.

Oppure possiamo fare come gli astronomi babilonesi: cerchiamo di raccogliere dati sui voli passati, sul costo dei biglietti nei diversi giorni prima della partenza, sul numero di posti già prenotati in occasione degli acquisti dei biglietti, su *tutto quanto possa influenzare il prezzo di un biglietto*, e cerchiamo di trovare delle correlazioni, che ci permettano di formulare un modello predittivo.

Oren Etzioni per primo ha cercato di *produrre un* modello predittivo per questo problema; inizialmente lavorò su dati relativi a 100.000 voli aerei, ma trovò che questi dati non erano sufficienti per arrivare a un metodo predittivo affidabile. Successivamente confrontò circa 200 miliardi di voli aerei per cui era noto il costo, il giorno di partenza, il giorno di acquisto del biglietto, gli aeroporti di partenza e arrivo, la compagnia aerea che effettuava il viaggio ed altri dati; e riuscì a produrre un modello affidabile e utile. Etzioni affermò che i viaggiatori che usarono il suo modello predittivo hanno in media risparmiato 50 dollari per ogni viaggio aereo, rispetto a quando decidevano per conto proprio.

In Figura 2.9 vediamo un terzo caso. Vediamo in figura quattro pneumatici dotati di sensori; nella figura c'è un solo sensore per pneumatico, ma nella realtà gli pneumatici sono dotati di decine di sensori che a seconda di dove vengono disposti misurano il grado di consumo del

pneumatico, il tipo di asfalto della strada percorsa, il sobbalzo del pneumatico e molti altri parametri. La loro analisi mediante un modello predittivo effettuata su un gran numero di pneumaici permette di valutare quando convenga sostituire un pneumatico; questa attività viene anche chiamata *manutenzione predittiva*.

Figura 2.9 – Quando conviene sostituire un pneumatico?

Torneremo diffusamente nel libro sui modelli predittivi, ed in particolare nella discussione sull'esempio Compas. Possiamo dire che i modelli predittivi ci dicono *cosa accadrà in futuro*. I modelli classificatori e predittivi sono quelli in cui prevalentemente vengono applicate le tecniche di Machine learning, e sono al centro del nostro interesse nel seguito del libro.

2.5 Modelli prescrittivi o decisionali

I modelli *prescrittivi* o decisionali ci aiutano a prendere una decisione riguardo a un problema con molte soluzioni diverse, soppesando e confrontando i vantaggi connessi alle diverse soluzioni.

Nella nostra vita, in ogni momento dobbiamo prendere decisioni, alcune volte semplici, vedi Figura 2.10, altre volte delle vere e proprie scelte di vita.

Figura 2.10 – Andiamo a sinistra o a destra?

Quando per il nostro lavoro o nella nostra vita privata dobbiamo prendere una decisione, sempre più spesso veniamo assistiti da un modello decisionale automatico. Per esempio, se siamo in automobile nei pressi di Porta Pia a Roma, e dobbiamo andare a San Pietro, possiamo consultare il nostro navigatore, oppure l'applicazione Waze, che ci forniscono alcuni percorsi alternativi con il tempo previsto nel caso si segua quel percorso, vedi Figura 2.11.

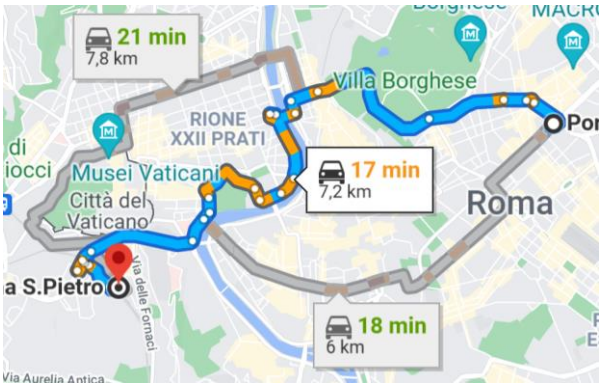


Figura 2.11 – Un navigatore che per un tragitto presenta due soluzioni

Il modello decisionale del navigatore deve prima costruire i percorsi nel grafo delle strade di Roma che vanno da Porta Pia a San Pietro, e poi deve aggiungere i tempi necessari per percorrere le varie tratte. Questa informazione, i tempi di percorrenza in quell'ora, possono essere fornite in tempo reale al modello calcolando la intensità del traffico sulla base del numero dei telefoni cellulari attivi nelle auto che stanno transitando nelle strade; oppure possono essere ricavate da un modello predittivo, in grado di prevedere il traffico sulla base dei dati storici.

Nel caso precedente, il modello decisionale propone diverse soluzioni al guidatore, e questi può decidere quale preferisce. In altre circostanze, come lo studio di caso Compas per la concessione della libertà provvisoria, la decisione si basa sulla previsione effettuata dal modello predittivo in merito al rischio di recidiva.

La previsione sul livello di rischio di recidiva può avere come esiti, ad esempio, rischio molto alto, alto, medio, basso, molto basso. A questo punto, la decisone può essere assunta *automaticamente* stabilendo una corrispondenza tra i differenti livelli di rischio e la decisione sulla libertà provvisoria concessa, vedi ad esempio Figura 2.12.

Rischio	**Decisione**
Molto alto	Libertà negata
Alto	Libertà negata
Medio	Libertà concessa
Basso	Libertà concessa
Molto basso	Libertà concessa

Figura 2.12 – Previsioni sul rischio e decisioni sulla libertà: una possibile corrispondenza

Nel seguito noi prenderemo in considerazione i modelli decisionali in due contesti: o come modelli considerati autonomamente, ovvero come modelli usati *insieme ai* modelli classificatori e predittivi. Quando studieremo i modelli decisionali insieme ai modelli predittivi, vale la seguente separazione dei loro ruoli:

- i modelli predittivi saranno prodotti mediante tecniche di Machine learning
- i modelli decisionali saranno, come nel caso Compas, *funzionalmente legati* ai modelli predittivi che forniscono loro gli elementi per la decisione, come accade in Figura 2.12.

Come conseguenza, i concetti di *accuratezza e equità saranno riferiti ai modelli predittivi*, ma, visto il legame funzionale tra modelli predittivi e modelli decisionali, faranno riferimento sia agli uni che agli altri.

Conclusioni

Per descrivere il mondo rappresentato dai dati digitali, ed elaborare a partire da essi conoscenza utile a classificare, predire, decidere, abbiamo a disposizione diverse tipologie di modelli. In molti casi questi modelli sono prodotti nell'ambito di scienze come la Fisica, la Matematica, la Statistica che nel corso dei secoli hanno investigato i fenomeni fisici (la Fisica), i fenomeni quantitativi (la Matematica), i fenomeni caratterizzati da incertezza (la Statistica).

Molti problemi possono ora essere affrontati e risolti grazie alle grandi quantità di dati digitali a nostra disposizione; ad esempio, il calcolo del giorno più conveniente per acquistare un biglietto aereo è stato possibile tramite un modello predittivo solo avendo a disposizione dati su miliardi di voli.

I modelli *classificatori* permettono di individuare il valore ignoto e non misurabile che assume una caratteristica del mondo attorno a noi. Questo valore è stimato dal modello partendo da un insieme di altre caratteristiche i cui valori conosciamo.

Nei modelli *predittivi*, la caratteristica cui siamo interessati è proiettata nel futuro, ad esempio il prezzo che avrà un biglietto aereo in un certo giorno davanti a noi.

I modelli *decisionali* ci aiutano a prendere una decisione tra diverse alternative.

Per quanto riguarda i *modelli classificatori e i modelli predittivi, quando ne parleremo in uno specifico contesto, useremo entrambe le dizioni; quando ne parleremo in generale, useremo il termine di modello predittivo, perché più immediatamente comprensibile e vivido rispetto al termine di modello classificatorio, che è un po' troppo astratto.*

Quando considereremo i modelli decisionali in connessione con i modelli predittivi, assumeremo che le decisioni formulate siano funzionalmente collegate con le previsioni del modello predittivo, ad esempio a rischio alto corrisponde sempre libertà provvisoria negata, a rischio basso sempre libertà provvisoria concessa.

Per produrre i modelli classificatori e predittivi, sono disponibili *tecniche,* molte delle quali si collocano nell'ambito della disciplina chiamata Machine learning. L'uso di queste tecniche per

la produzione di modelli classificatori e predittivi pone problemi etici rilevanti, come stiamo per scoprire.

Quindi (scusate la ripetizione):

Un modello classificatorio o predittivo ci permette di assegnare un valore a una caratteristica y di un fenomeno che non possiamo misurare o conoscere direttamente.

Una tecnica di apprendimento (o più semplicemente tecnica) ci permette di costruire un modello classificatorio o predittivo tramite un processo di apprendimento basato su dati del passato.

Modelli classificatori e predittivi e tecniche di apprendimento sono i concetti centrali che discuteremo nel resto del libro.

Tutto chiaro?

Sì, tutto chiaro; se posso esprimere una critica, sei un po' ripetitivo…

Va bene, cercherò di essere più stringato, ma certe volte repetita iuvant..

Capitolo 3. Le tecniche di Machine learning

Il Machine learning è la disciplina che studia le tecniche basate su apprendimento che permettono di costruire modelli classificatori e predittivi. Le tecniche utilizzano metodi basati sulla statistica, il calcolo della probabilità, la logica, l'informatica, e vanno dagli alberi di decisione, alle foreste di alberi, alle reti neurali.

Le tipologie di Machine learning più studiate e applicate sono sviluppate nel seguito.

3.1 Machine learning da esempi o supervisionato

Questo è il caso che ho introdotto nelle Figura 1.6 e Figura 1.7 commentando il caso Compas. Nel Machine learning da esempi sono forniti alla tecnica di apprendimento, come esempio, dati di input e rispettivi dati di output ad essi collegati tramite una funzione che associa a ciascun input il corrispondente output; per quei dati di esempio la funzione tra dati di input e dati di output è nota.

Nel caso Compas la funzione tra input e output è quella che fa corrispondere alle caratteristiche di ciascun detenuto: a. genere, b. reato contestato, c. numero di arresti precedenti, l'aver commesso o meno recidiva:

[genere, reato contestato, numero di arresti] → Recidiva?

Il modello predittivo viene costruito utilizzando un albero di decisione e contando per ogni foglia dell'albero il numero di detenuti che hanno e non hanno commesso recidiva. Ogni foglia corrisponde a una terza di valori delle tre caratteristiche, ad esempio [uomo, reato grave, nessun arresto]. A seconda del valore prevalente *viene assegnato a tutti i detenuti della foglia* rischio alto o rischio basso.

A questo punto, abbiamo creato il modello predittivo; quando un nuovo detenuto fa domanda di libertà provvisoria, si individua la foglia a cui corrispondono le sue caratteristiche, e si associa il rischio associato alla foglia. Il processo che porta a costruire un modello predittivo è chiamato *processo di classificazione*.

Il Machine learnig da esempi è applicato in moltissime aree. Ad esempio, per la concessione di prestiti ai clienti, le banche devono determinare il livello di rischio che il cliente non restituisca il prestito, chiamato rischio di credito. Per valutarlo, viene creato un modello predittivo basato su una tecnica di apprendimento che associa ad ogni cliente del passato un valore [sì, ha restituito il prestito; no, non ha restituito il prestito]. Se la tecnica è un albero di decisione, viene associato a ogni foglia dell'albero il valore prevalente tra il sì e il no. Per i nuovi clienti che chiedono un prestito, il livello di rischio associato è quello della foglia a cui corrisponde il cliente.

Tutte le precedenti attività sono discusse nei capitoli successivi, il libro è sostanzialmente dedicato alle proprietà dei modelli predittivi costruiti a partire da tecniche di apprendimento basate su esempi.

3.2 Machine learning ensemble

Il Machine learning ensemble è un Machine learning da esempi in cui, invece di costruire un solo modello predittivo, se ne costruiscono tanti, a ciascuno dei quali sono forniti dati di ingresso che derivano da un unico insieme, e che contengono solo alcune righe e colonne dei dati originari, vedi Figura 3.1.

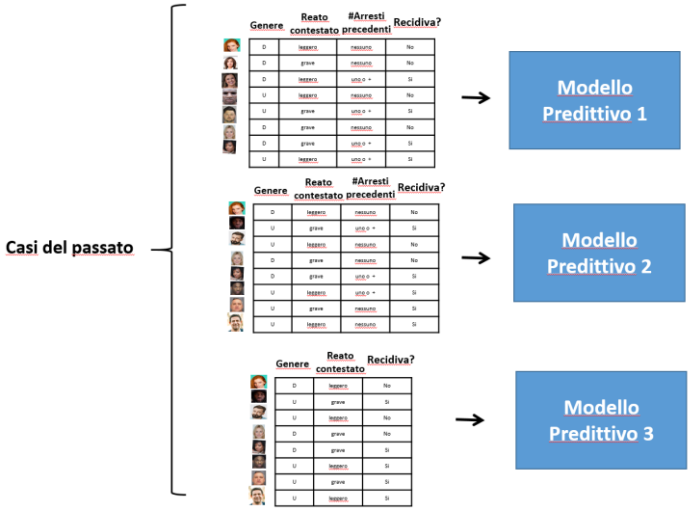


Figura 3.1 – Esempio di Machine learning ensemble

Il modello viene a questo punto costruito mediante un meccanismo di votazione. Ad esempio, se tre modelli hanno rischio alto per una foglia, e cinque hanno rischio basso, nel modello finale la foglia ha associato rischio basso, vedi Figura 3.2. I voti possono anche essere pesati, ad esempio il voto di un modello può valere di più perché il modello è valutato più accurato.

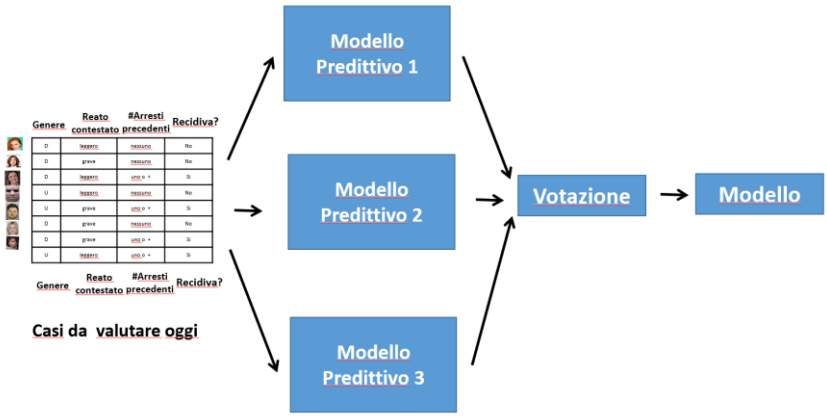


Figura 3.2 – Costruzione del modello per votazione

3.3 Machine learning non supervisionato

In questo caso alla tecnica di apprendimento sono forniti, come esempio, dati di input senza che questi siano associati a valori di output; la tecnica ha il compito di riconoscere schemi/strutture/proprietà nell'insieme dei dati fornito. Per esempio, in Figura 3.3 vediamo un insieme di libri di una biblioteca; la tecnica di learning non supervisionato ha il compito di raggruppare (clustering) i libri in due categorie, libri di saggistica e romanzi, vedi Figura 3.4. Poiché

la tecnica non dispone di esempi di libri classificati nelle due categorie, dovrà utilizzare altre proprietà e caratteristiche dei libri per procedere nella classificazione.

Figura 3.3 – I libri di una biblioteca

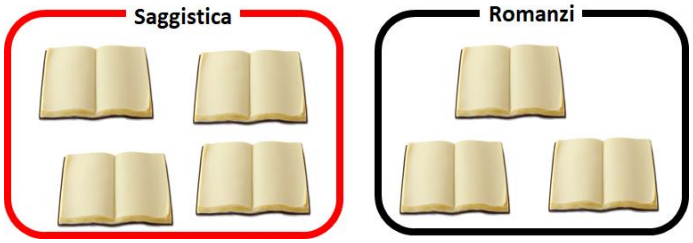


Figura 3.4 – I libri classificati in libri di saggistica e romanzi

In Figura 3.5. vediamo un insieme di prelievi bancari effettuati nel tempo, che supponiamo essere di tre differenti somme, 500 euro, 1.000 euro e 2.000 euro. In questo caso la tecnica di learning deve diagnosticare prelievi anomali (anomaly detection).

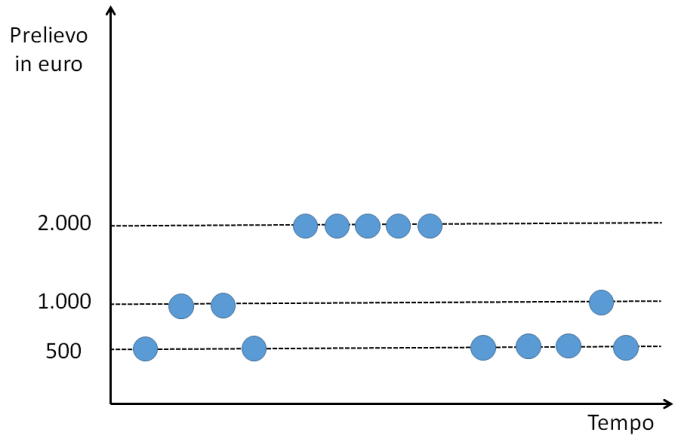


Figura 3.5. – Esempio di ricerca di anomalie

Sulla base dell'ammontare dei prelievi e della loro densità nel tempo, possiamo arrivare alla conclusione che i cinque prelievi consecutivi di Figura 3.6 sono anomali, perché sono gli unici di 2.000 euro e perché sono consecutivi e vicini nell'arco temporale.

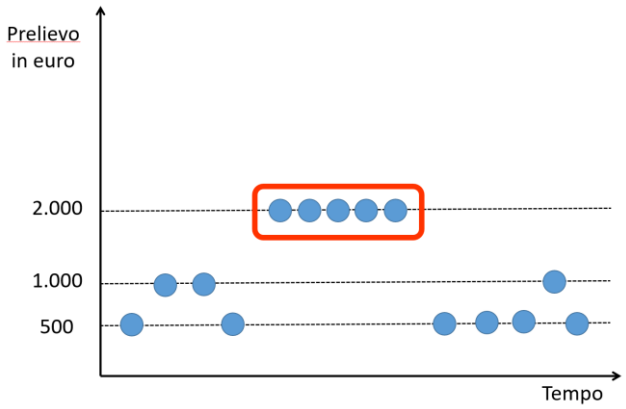


Figura 3.6 – Serie di prelievi anomali

Non approfondiremo nel resto del libro il Machine learning non supervisionato.

3.4. Machine learning per rinforzo

In questo caso il modello predittivo viene generato mediante apprendimento sulla base delle risposte dell'ambiente alle sue azioni e decisioni; ad esempio, se lo scopo del modello è giocare a scacchi, vengono fatte giocare al modello tante partite in cui le singole mosse e le sequenze di mosse vengono premiate o penalizzate sulla base dell'esito della partita, vittoria o sconfitta.

Possono essere generati diversi modelli che giocano tra di loro, così che alla fine prevalga quello che presenta le migliori prestazioni. Le tecniche di apprendimento tipicamente usate nel Machine learning per rinforzo sono le reti neurali.

AlphaZero è un modello per il gioco degli scacchi ed altri giochi come Go, basato su tecniche di Machine learning per rinforzo e sviluppato da Google DeepMind nel 2017, che poco tempo dopo la sua produzione ha battuto in una sfida di 100 partire l'algoritmo campione del mondo di scacchi Stockfish.

Altri esempi di applicazioni del Machine learning per rinforzo sono quelli che consentono ad un robot di muoversi all'interno di un ambiente con un certo fine.

Conclusioni

Abbiamo visto che il Machine learning è una disciplina che corrisponde a diverse "filosofie", può basarsi su esempi che forniscono la funzione obiettivo (ha commesso recidiva?), può basarsi su una esplorazione libera, non aiutata da esempi, può basarsi sul meccanismo del premio o della punizione. Tutti questi meccanismi, in fondo, corrispondono a strategie che utilizziamo nella vita di ogni giorno, niente di nuovo sotto il sole.

Capitolo 4 - Come funziona una tecnica di Machine learning

In questo capitolo vediamo all'opera le tecniche di apprendimento basate su esempi, concentrando l'attenzione su quella più semplice e intuitiva, gli alberi di decisione.

4.1 Prevedere e decidere

In questo capitolo ci concentriamo sul caso Compas. Riassumiamo quanto abbiamo appreso fino a qui. La decisione che deve prendere un giudice quando un detenuto fa richiesta di libertà provvisoria è tra le più delicate, per le conseguenze sociali e psichiche che la detenzione e la mancata liberazione provocano nei detenuti in attesa di giudizio; questi detenuti, come sappiamo, devono essere considerati innocenti di fronte alla legge fino a giudizio definitivo.

Negli Stati Uniti da tempo sono usati diversi modelli predittivi che aiutano i giudici nel prendere questo tipo di decisioni; uno strumento tra i più utilizzati è lo strumento Compas, vedi Figura 4.1.

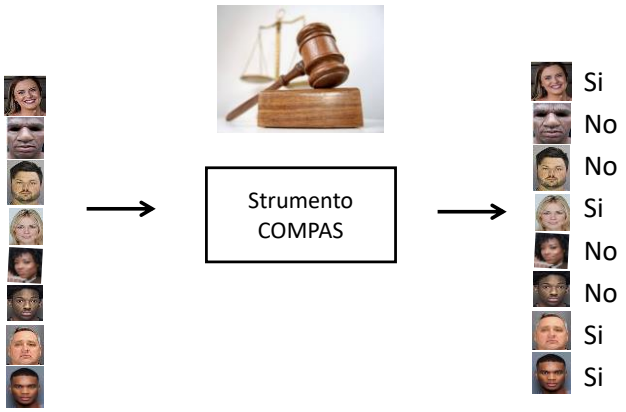


Figura 4.1 - Lo strumento Compas utilizzato negli Stati Uniti

Compas acquisisce i profili dei detenuti che hanno fatto richiesta di libertà provvisoria, e restituisce in output il *rischio che il detenuto possa commettere un nuovo reato mentre è in libertà*. A seconda del livello di rischio, e, in alcuni casi, altri elementi raccolti, il giudice prende la decisione definitiva: libertà concessa o libertà negata, rappresentate dai sì e i no nella Figura 4.1.

Compas può essere visto (vedi Figura 4.2) come un modello predittivo che fornisce al giudice il livello di rischio di recidiva, espresso in una scala o di più valori (ad esempio: rischio alto, rischio medio, rischio basso); il giudice, ovvero un semplice modello decisionale, sulla base di una soglia tra i livelli di rischio (ad esempio, nella scala a tre valori: a rischio alto corrisponde libertà negata, a rischio medio e basso corrisponde libertà concessa) stabilisce se concedere o meno la libertà, vedi Figura 4.2.

Noi nel seguito assumeremo che il livello di rischio sia a due valori, rischio alto e rischio basso, e che a ciascuno dei due corrisponda una decisione, libertà negata o libertà concessa.

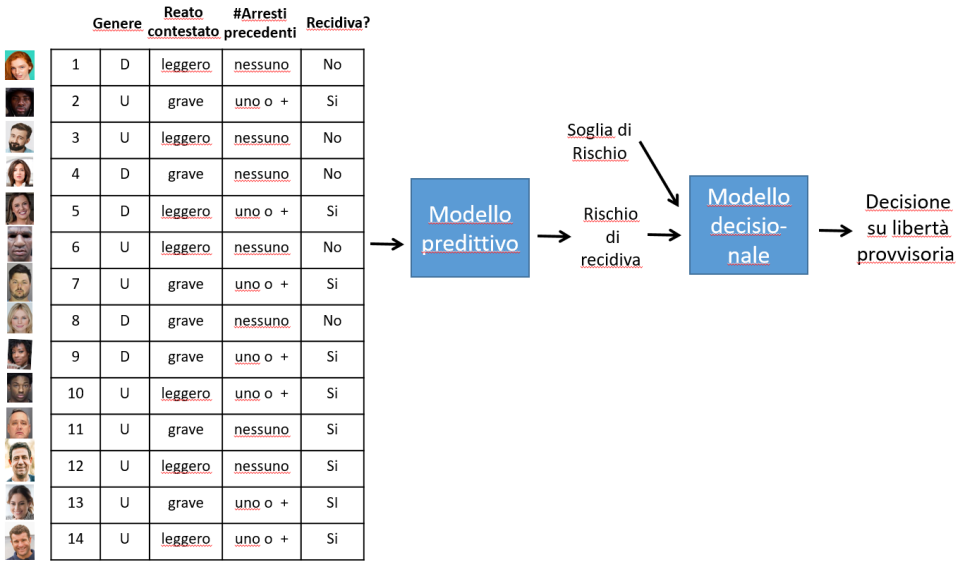


Figura 4.2 – Compas come modello predittivo

Compas si basa su un'idea semplice, quella di prendere in considerazione, a partire dall'universo dei detenuti a cui nel passato è stata concessa la libertà provvisoria, un *campione* di detenuti, vedi Figura 4.3; per tali detenuti vengono registrate:

- alcune caratteristiche anagrafiche e sulla storia pregressa del detenuto, in questo caso semplificato il genere.
- il tipo di reato contestato per cui il detenuto si trovava in carcere, se grave, ad esempio un omicidio, ovvero leggero, ad esempio un furto di merce di trascurabile valore economico, e
- il numero di eventuali arresti precedenti all'ultimo.
- se abbiano o meno commesso successivamente recidiva.

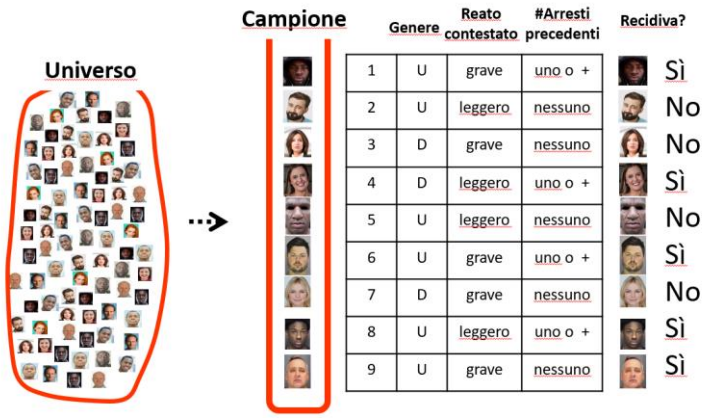


Figura 4.3 – Cosa è accaduto nel passato?

Fermiamoci un attimo, perché nella nostra narrazione è già accaduto qualcosa di molto importante, che avrà grande influenza nel seguito del discorso. Tutti gli operatori della società civile, quando prendono una decisione che ha un peso sulla comunità delle persone, possono assumere come loro riferimento:

- *il mondo ideale,* cioè *il mondo come dovrebbe o potrebbe essere*, ovvero, più conservativamente,
- *il mondo reale,* cioè *il mondo così come è*, *e come è stato nel passato*, vedi Figura 4.4.

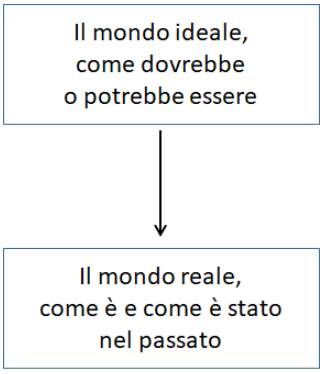


Figura 4.4 – Mondo ideale e mondo reale

Un giudice è guidato nelle sue decisioni dal codice di procedura penale, ma può tener conto della storia pregressa del detenuto, del contesto familiare in cui ha vissuto; questi aspetti possono portarlo a decidere di concedere la libertà provvisoria, anche quando i dati disponibili sul detenuto lo configurano come detenuto ad alto rischio di recidiva. Se il giudice tiene conto di elementi di riabilitazione basati sulla ricerca nelle scienze sociali o in psicologia, guarda al mondo come dovrebbe o potrebbe essere, più che al mondo come è oggi, vedi ancora Figura 4.4.

Nel caso il giudice decida di fare uso di Compas, il punto di vista che assume è semplicemente di *guardare al passato*, senza nessun tentativo di modificare la natura sociale del fenomeno della libertà provvisoria in futuro; osserva semplicemente il mondo come è e come è stato, vedi ancora Figura 4.4.

Facciamo ora un passo in avanti: come fa Compas a utilizzare *le informazioni sul passato* per decidere su quanto fare oggi? Troviamo sulla nostra strada il concetto di *somiglianza*. Compas non decide partendo dalla persona, dalla sua unicità e identità nel mondo, ma sulla base di *quanto quella persona è simile* alle persone che sono state liberate nel passato.

Naturalmente, la somiglianza può essere calcolata in modo efficace solo se si hanno a disposizione un insieme ragionevolmente ampio di *caratteristiche comuni* per le persone del passato e quelle del presente. In Figura 4.5 vediamo un estratto delle caratteristiche conservate nell'archivio storico dei detenuti dello stato della Pennsylvania. Tra esse, vediamo il genere, l'età, il reato attuale per cui si è detenuti, il numero di arresti precedenti.

Scale	OGS 1 n=6673 0-10	OGS 2 n=5687 0-10	OGS 3 n=18021 0-16	OGS 4 n=2328 0-9	OGS 5 n=6946 0-13	OGS 6 n=4126 0-9	OGS 7 n=2599 0-9	OGS 8 n=1140 0-6	OGS 9-14 n=3221 0-8
Gender		Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0
County	Alleg =1 ll other=0	Urban=1 Rural=0	Alleg =1 All other=0		Urban=1 Rural=0				
Age	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21 to 29=2 30-44=1 >44=0	<21=3 21-25=2 26-39=1 >39=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=2 21 to 39=1 over 39-0	<21=3 21-29=2 30-49=1 >49=0
Current offense			Property Fel=1 All other=0						
Number of Prior Arrests	none=0 1=1 2 to 4 =2 5 to 9 =3 over 9=4	none=0 1=1 2=2 3 to 6=3 over 6=4	none=0 1=1 2=2 3 to 4=3 5 to 7=4 over 7=5	none=0 1 to 2=1 3 to 8=2 over 8=3	none=0 1=1 2 to 4=2 5 to 7=3 over 7=4	none, 1=0 2=1 3 to 6=2 over 6=3	none=0 1=1 2 to 6=2 over 6=3	none=0 1 to 4=1 over 4=2	0=0 1=1 2 to 4=2 5 to 7=3 Over 7=4
Prior Offense Type	c order=1 drug=1	drug=1	property=1 drug=1 public adm.=1	drug=1	drug=1 public adm=1				
Multiple charges			Yes=1 No =0	Yes=1 No =0	Yes=1 No =0	Yes=1 No =0	Yes=1 No =0		
PRS							Yes=1 No =0	Yes=1 No =0	
Prior juv. Adjud			Yes=1 No/ unknown=0		Yes=1 No/ unknown=0	Yes=1 No/ unknown=0			

33

Figura 4.5 – Caratteristiche conservate nell'archivio storico dei detenuti della Pennsylvania

Nel nostro studio di caso, assumeremo che i precedenti su cui basare il processo di apprendimento siano quelli di Figura 4.6. *Noi assumiamo qui e nel seguito di prendere in considerazione solo detenuti a cui è stata concessa nel passato la libertà provvisoria*.

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	D	grave	nessuno	No
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	uno o +	Si
8	D	grave	nessuno	No
9	D	grave	uno o +	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	leggero	nessuno	Si
13	U	grave	uno o +	No
14	U	leggero	uno o +	Si

Figura 4.6 - I dati disponibili su quanto è accaduto nel passato

Stiamo parlando di 14 detenuti, di cui supponiamo di conoscere le caratteristiche già viste in precedenza. Se guardate la figura 4.6, arriviamo a un'altra conclusione nuova e importante: come evidenziato in Figura 4.7, noi non solo con Compas decidiamo il futuro guardando soltanto al passato, ma guardiamo al passato attraverso occhiali che osservano soltanto *genere, reato contestato e numero di arresti precedenti*, informazioni certamente utili, ma molto limitate nel rappresentare persone come i 14 della Figura 5.6.

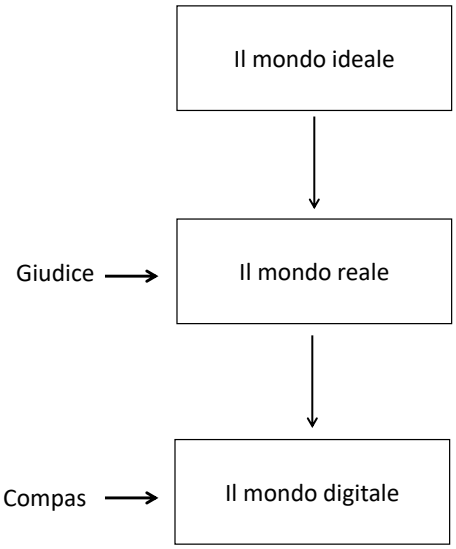


Figura 4.7 – Compas non prende in considerazione il mondo,
ma una limitata porzione del mondo

E' ben vero che usualmente i modelli basati sull'apprendimento automatico utilizzano decine e centinaia di caratteristiche, riguardanti un universo molto più densamente popolato di quello del nostro esempio, ma il principio è sempre lo stesso: fotografare realtà complesse, come nel nostro caso gli esseri umani, attraverso pochi caratteri; è come fotografare un panorama con pochi pixel, verrà una foto molto sgranata e incompleta, senza sfumature e con scarsa comprensione della complessità che caratterizza gli esseri umani.

Torniamo ora ai 14 casi della Figura 4.6, e cerchiamo di capire se esistano delle regole ricorrenti nei comportamenti dei detenuti in libertà provvisoria, costruendo la tabella di Figura 4.8. Nella tabella, per ogni combinazione possibile dei valori di genere, reato contestato e numero di arresti precedenti, calcoliamo i casi in cui è stata o non è stata commessa recidiva.

Genere	Reato contestato	# Arresti precedenti	No-recidiva	Si-recidiva
D	leggero	nessuno	1	0
D	leggero	uno o +	0	1
D	grave	nessuno	2	0
D	grave	uno o +	0	1
U	leggero	nessuno	2	1
U	leggero	uno o +	0	2
U	grave	nessuno	0	1
U	grave	uno o +	1	2

Figura 4.8 – Gli otto casi possibili, e i relativi comportamenti nel passato
delle persone in libertà provvisoria

La tabella di Figura 4.8 ci fornisce una rappresentazione molto "piatta" delle diverse storie connesse ai 14 casi di Figura 4.6; noi possiamo rappresentare la figura in un altro modo, mediante un albero, come in Figura 4.9: vediamo come costruirlo.

Per l'insieme delle persone descritte in Figura 4.6 possiamo chiederci: è donna o uomo? Se è donna percorriamo il ramo sinistro dell'albero di Figura 4.9, se uomo il ramo destro. Sia per le donne che per gli uomini calcoliamo quanti non hanno commesso recidiva e quanti l'hanno commessa e riportiamo i due dati vicino al ramo; ad esempio la notazione <2;5> accanto al ramo associato agli uomini significa che:

- due uomini non hanno commesso recidiva
- cinque uomini hanno commesso recidiva.

L'albero così costruito è mostrato in Figura 4.9, in cui l'albero è ad un solo livello.

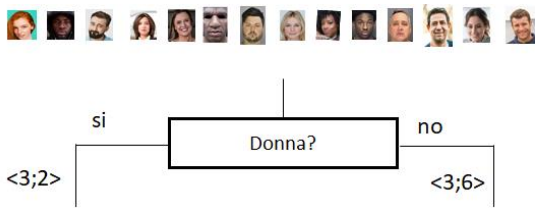


Figura 4.9 – Un albero di decisione ad un livello

A questo punto, proseguiamo nella creazione dell'albero utilizzando dapprima la caratteristica *Reato contestato* e poi la caratteristica *Numero di Arresti precedenti*, riportando sempre per ogni ramo quanti non hanno commesso recidiva e quanti l'hanno commessa. Otteniamo così l'albero di Figura 4.10, in cui in corrispondenza dei cammini che non si diramano ulteriormente (le *foglie dell'albero*) riportiamo il numero delle non recidive e delle recidive. *Notate un aspetto importante: qui e nel seguito del libro ogni caso, ogni detenuto, vale uno, tutti i casi hanno lo stesso peso, pari a uno.*

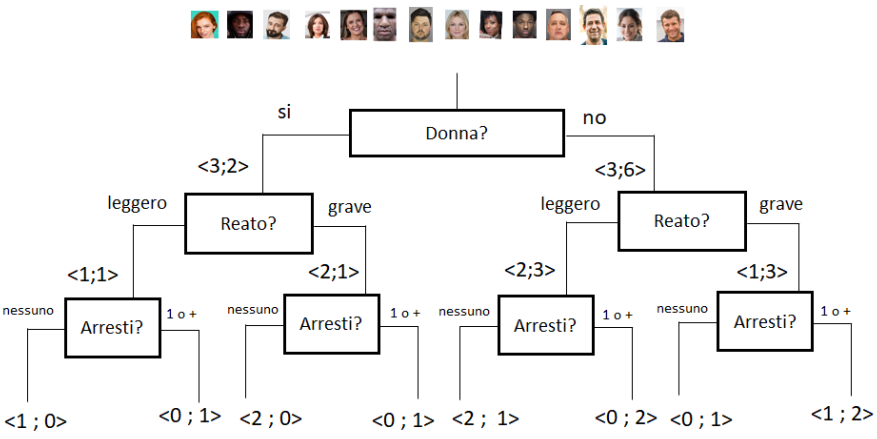


Figura 4.10 - Albero dei casi possibili e relative recidive e non recidive

Vi propongo ora un esercizio, che, come i successivi, può essere saltato dal lettore non interessato a esercitarsi.

Esercizio 4.1 - Partendo dalle 14 persone di Figura 4.6, rappresentate ciascuna con un cerchietto con i numeri 1, 2, …, 14, e poi associate i diversi cerchietti alle foglie della Figura 4.10, a sinistra se la persona non ha commesso recidiva, a destra se la ha commessa. La risposta nella prossima pagina.

Per assegnare i cerchietti alle foglie, occorre per ogni persona seguire il percorso associato alle tre caratteristiche della persona, e associarlo al valore a sinistra o a destra a seconda che non abbia commesso o abbia commesso recidiva.

Risposta all'esercizio 4.1.

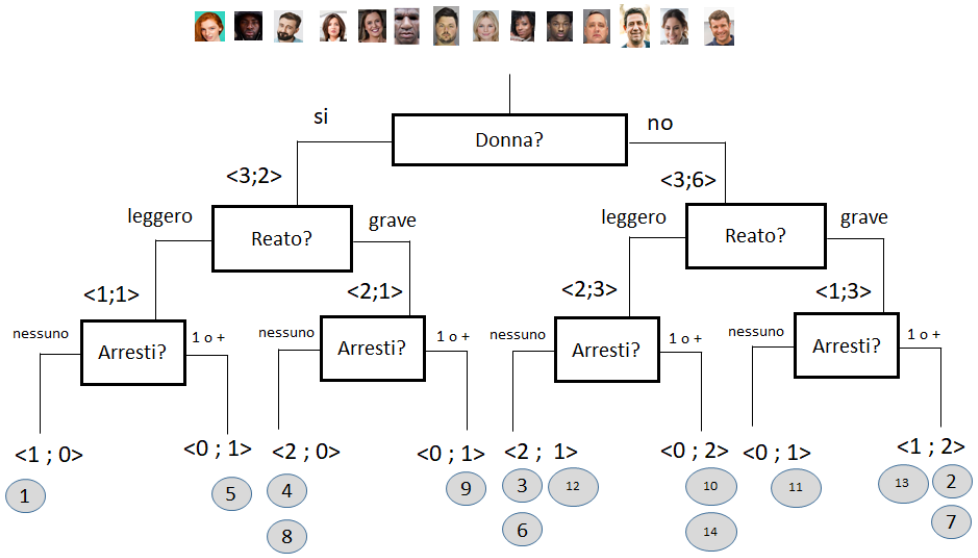


Figura 4.11 – Risposta all'esercizio

Abbiamo dunque costruito un albero, che chiameremo *albero di decisione.* L'albero di decisione suddivide i casi del passato per tutte le combinazioni possibili. A questo punto, per produrre un albero che possa essere utilizzato per le nuove richieste di libertà provvisoria ci rimane un ultimo passo: *per ogni foglia decidiamo*, sulla base del valore prevalente tra non recidiva e recidiva, il livello di rischio delle persone associate alla foglia, vedi Figura 4.12.

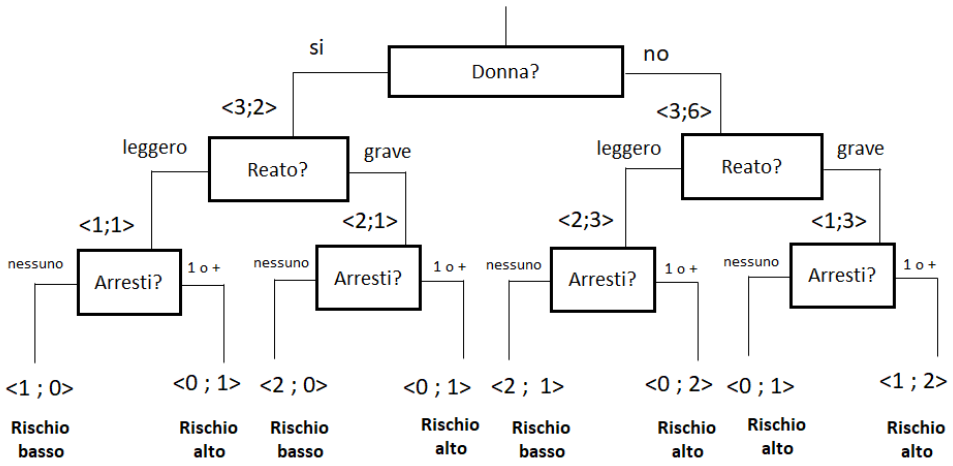


Figura 4.12 – Albero dei casi possibili e livelli di rischio per i detenuti

Notate che in tutti i casi eccetto i casi <2;1> e <1;2> i nodi foglia fotografano situazioni chiare, o tutte le persone hanno commesso recidiva (associamo *rischio alto*) o non la hanno commessa (associamo *rischio basso*). Quindi, ad esempio, per la foglia più a sinistra, sulla base dei due valori <1;0> decidiamo per rischio basso, per la foglia più a destra sulla base dei due valori <0;2>, decidiamo per rischio alto. Nel caso <2;1> decidiamo per rischio basso secondo un criterio di prevalenza. Attenzione: questo comporta il fatto che, ad esempio, alla persona

13 di Figura 4.11, pur non avendo commesso recidiva, viene assegnato rischio alto; questi casi saranno chiamati d'ora in poi falsi positivi, perché risultano avere falsamente rischio alto.

Abbiamo in questo modo costruito un modello predittivo (vedi Figura 4.13) che possiamo utilizzare per i nuovi detenuti; per ciascuno, basta percorre l'albero fino alla foglia che corrisponde alle sue caratteristiche, attribuendo il livello di rischio relativo.

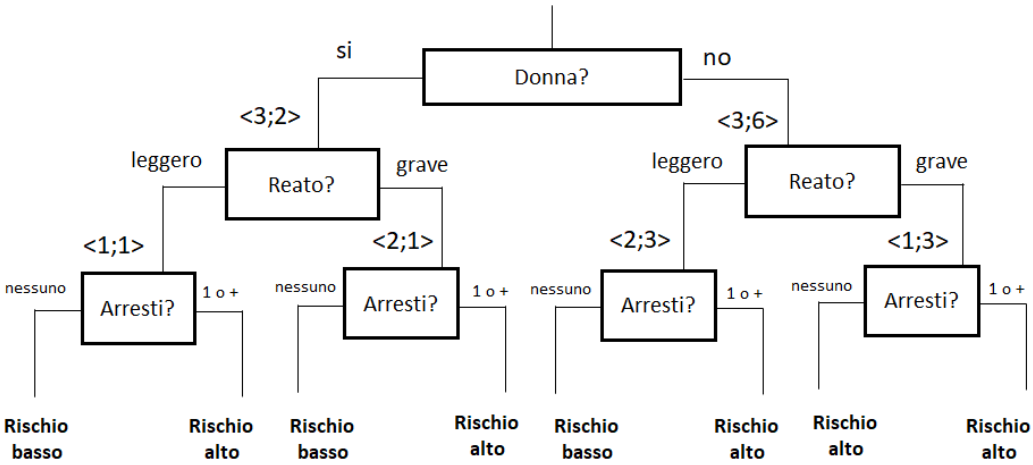


Figura 4.13 – Modello predittivo

Occorre ora osservare che una scala di rischio a due valori ha senso quando si parte da un numero limitato di esempi. Immaginate un numero molto elevato di osservazioni, ad esempio 10.000; in questo caso potranno verificarsi casi come <10; 100>, <50; 200>, <100; 250>, <300;200>, e così via. E' chiaro che qui potremo adottare una scala di rischio più ampia, per esempio con cinque livelli di rischio; oppure possiamo adottare una scala numerica, calcolando il rapporto tra valore a sinistra e somma dei due valori, ottenendo così per i casi precedenti 1, 0, 1, 0.66, 0, 0.33. Potremmo a questo punto definite una soglia (ad esempio 0.6) che separerà due gruppi, quelli a cui concedere e quelli a cui negare la libertà provvisoria.

Ricapitolando: assumendo di associare a rischio basso la concessione della libertà provvisoria e a rischio alto la non concessione (vedi Figura 4.14), la decisione verà presa per ogni richiedente percorrendo, sulla base delle sue caratteristiche l'albero di Figura 4.13; quando si arriva a una foglia, viene osservato il livello di rischio associato a quella foglia e presa la decisione.

Livello di rischio	**Decisione**
Basso	Si-libertà provvisoria
Alto	No-libertà provvisoria

Figura 4.14 – Modello decisionale

E' tutto chiaro?

Sì, è abbastanza chiaro; devo fare una osservazione, ma ….prosegui, non è ancora il momento di intervenire, voglio vedere come va a finire….

Nella Figura 4.15 riporto nuovamente la sequenza dei passaggi nella costruzione del modello predittivo e del modello decisionale.

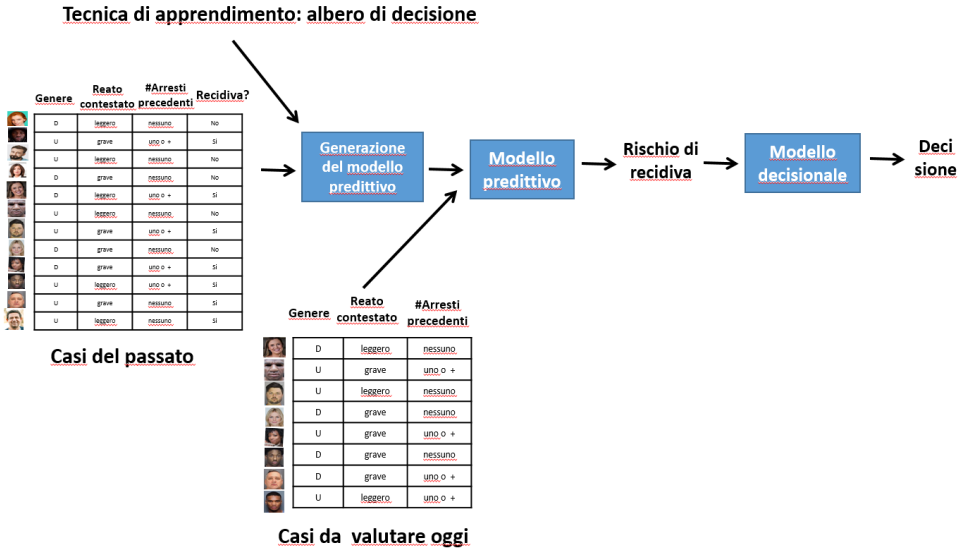


Figura 4.15 – Modelli predittivo e decisionale nel nostro caso

Adesso modifichiamo un po' l'esempio, cambiando in uno dei quattordici casi il valore di una caratteristica, vedi Figura 4.16.

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	D	grave	nessuno	No
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	uno o +	Si
8	D	grave	nessuno	No
9	D	grave	uno o +	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	leggero	nessuno	Si
13	U	grave	uno o +	No
14	U	leggero	uno o +	Si

→

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	**D**	**grave**	**uno o +**	**No**
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	uno o +	Si
8	D	grave	nessuno	No
9	D	grave	uno o +	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	leggero	nessuno	Si
13	U	grave	uno o +	No
14	U	leggero	uno o +	Si

Figura 4.16 – Cambiamo il valore di una caratteristica di uno dei quattordici casi

Esercizio 4.2 – Produci il nuovo albero di decisione.

In Figura 4.17 vediamo il nuovo albero di decisione. Cosa noti nella Figura?

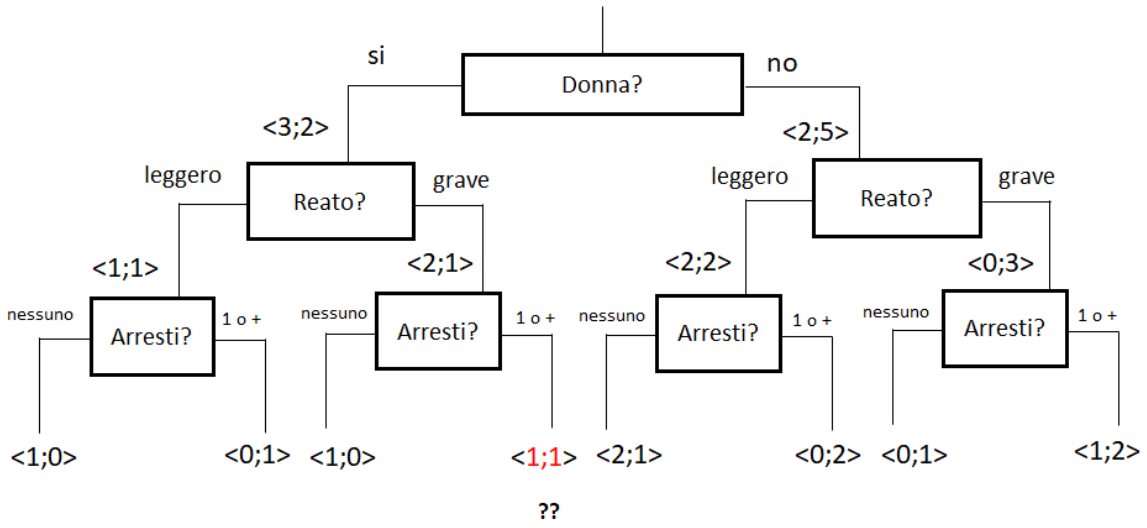


Figura 4.17 – Quale livello di rischio assumiamo per il caso
[donna, reato grave, 1 o + arresti +?

Noto che nel caso "donna; grave; 1 o + arresti" ora non so più decidere, perché i no i e si si equivalgono….

E' così; con esempi molto grandi questo caso è raro, ma ugualmente dobbiamo decidere cosa fare: tipicamente viene definito un valore detto di default, ad esempio si può convenire livello di rischio basso e libertà provvisoria concessa.

Senti, a questo punto ti ripeto quanto ti ho detto all'inizio: io provo un profondo disagio nel fare esercizi per una questione così delicata e così potenzialmente dolorosa in caso di rifiuto come la concessione della libertà provvisoria!

Ti capisco perfettamente, ma è proprio perché stiamo parlando di una questione molto dolorosa che vale la pena ragionare su questo caso, proprio perché il Machine learning è usato per queste decisioni che è necessario entrare nel profondo dei problemi, anche con esempi molto semplici che possono far sembrare il tutto un gioco sgradevole. Andiamo avanti…

Va bene…

Esercizio 4.3 - Per renderti conto se hai capito fino questo momento, ti propongo un esercizio, che riprenderemo anche nei capitoli successivi, quindi ti consiglio di svolgerlo. Guarda la Figura 4.18.

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	U	grave	nessuno	Si
2	D	leggero	nessuno	No
3	D	grave	uno o +	No
4	U	grave	uno o +	Si
5	U	leggero	nessuno	No
6	U	grave	nessuno	Si
7	D	grave	nessuno	No
8	D	leggero	uno o +	Si
9	U	leggero	uno o +	Si
10	U	leggero	nessuno	No
11	U	grave	uno o +	Si
12	D	grave	nessuno	No
13	D	grave	uno o +	Si
14	D	grave	nessuno	Si
15	U	leggero	uno o +	Si
16	U	grave	nessuno	Si
17	U	grave	uno o +	Si
18	U	leggero	nessuno	Si
19	D	leggero	nessuno	No
20	U	grave	uno o +	SI
21	D	grave	uno o +	No
22	U	leggero	nessuno	Si

Figura 4.18 – Un insieme di 22 persone da usare nell'Esercizio 4.3

Lo scopo dell'esercizio è:

Domanda 1 - Costruire le statistiche sulle recidive e non recidive per ciascuna delle otto combinazioni dei casi possibili per le tre caratteristiche, costruendo una tabella simile a quella di Figura 4.8.
Domanda 2 – Costruire l'albero di decisione.
Domanda 3 – Costruire il modello predittivo del rischio, attribuendo a ciascun delle foglie rischio alto o rischio basso a seconda del prevalere delle recidive o delle non recidive.
Domanda 4 – Costruire ora il modello decisionale, associando a ciascuna foglia dell'albero una delle due decisioni possibili, si-libertà provvisoria o no-libertà provvisoria.

Risposta alla Domanda 1.

Genere	Reato contestato	# Arresti precedenti	No-recidiva	Si-recidiva
D	leggero	nessuno	2	0
D	leggero	uno o +	0	1
D	grave	nessuno	2	1
D	grave	uno o +	2	1
U	leggero	nessuno	3	2
U	leggero	uno o +	0	2
U	grave	nessuno	0	3
U	grave	uno o +	0	3

Figura 4.19 – Risposta alla domanda 1

Risposta alla Domanda 2.

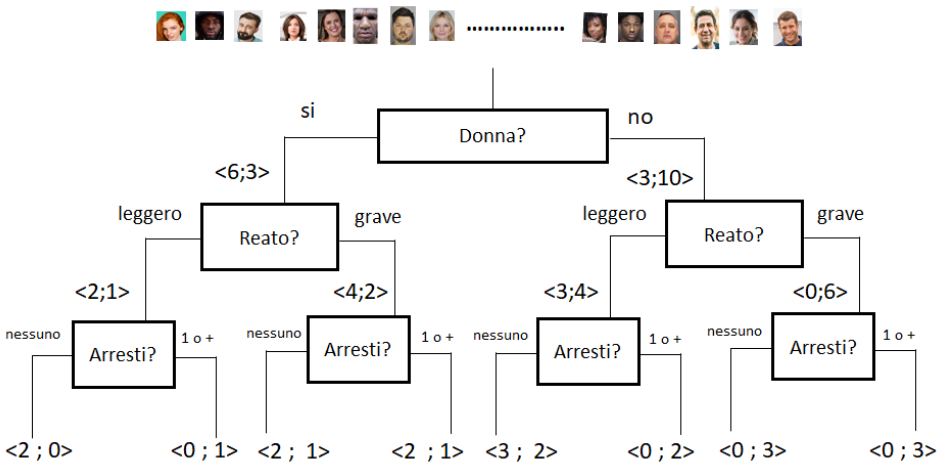


Figura 4.20 – Albero per la Domanda 2

Risposta alla Domanda 3.

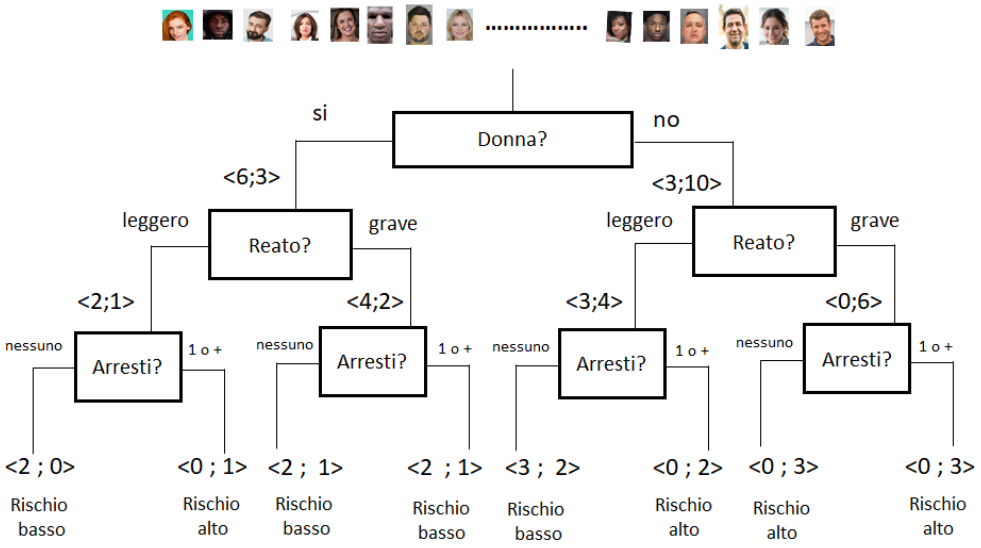


Figura 4.21 – Modello predittivo per la Domanda 3

Risposta alla Domanda 4.

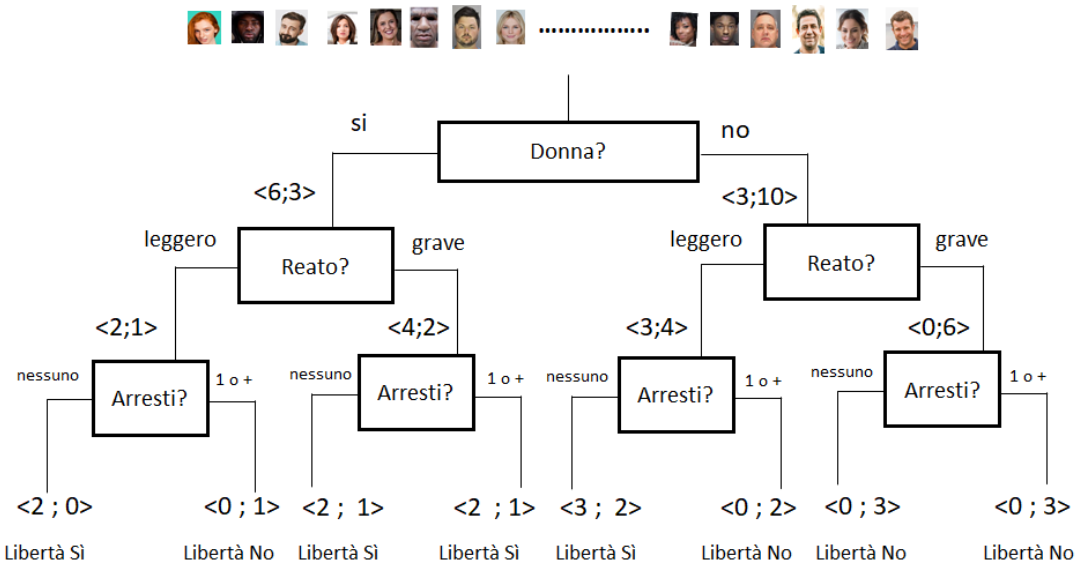


Figura 4.22 – Modello decisionale

4.2 L'entropia e il guadagno informativo

Mi piacerebbe che ci arrivassi da solo a capire i concetti di entropia e guadagno informativo. Ma attenzione, qui il percorso si fa più impervio. Per cui cominciamo con un esercizio.

Esercizio 4.4 – Considera la popolazione di Figura 4.23. Il nostro obiettivo è di capire per ciascuna delle tre caratteristiche quanto essa riesce a *separare* gli elementi della popolazione nei due insiemi con rischio basso (recidiva No) e alto (recidiva Sì).

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	D	grave	nessuno	No
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	uno o +	Si
8	D	grave	nessuno	No
9	D	grave	uno o +	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	leggero	nessuno	Si
13	U	grave	uno o +	SI
14	U	leggero	uno o +	Si

Figura 4.23 - La popolazione di partenza

Cosa intendi per separare?

Mi spiego con un esempio. In Figura 4.24 ho separato gli elementi della tabella di Figura 4.23 nei due insiemi corrispondenti agli elementi a basso rischio e ad alto rischio. In questo caso la separazione ha portato a separare i quattordici elementi in cinque a basso rischio e nove ad alto rischio.

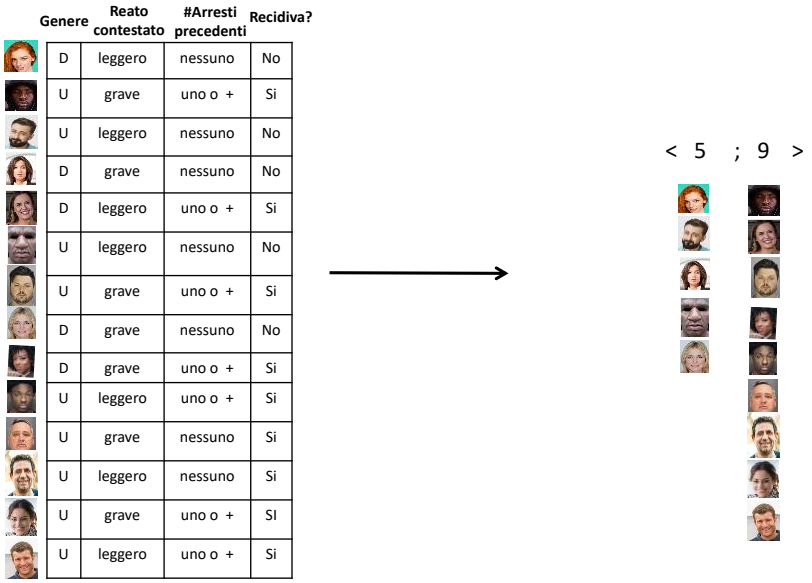


Figura 4.24 – Separazione degli elementi della popolazione tra elementi a basso rischio e ad alto rischio

Adesso genero l'albero di decisione per la caratteristica Genere; prima (Figura 4.25)

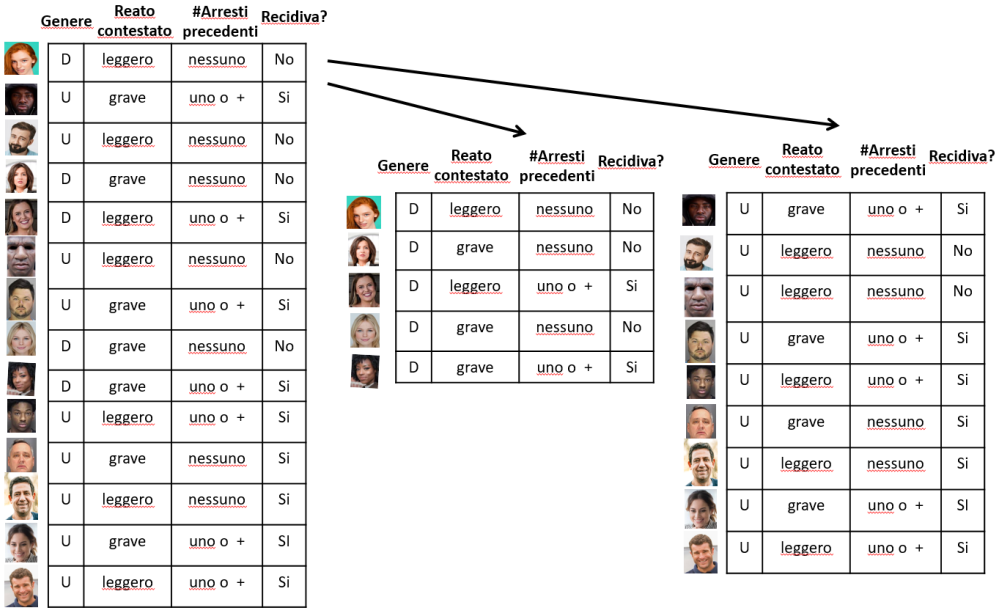


Figura 4.25 – Separazione della popolazione rispetto al genere

suddivido la tabella della popolazione iniziale in due tabelle, corrispondenti alle donne e agli uomini. Ora produco l'albero di decisione per la caratteristica Genere, vedi Figura 4.26.

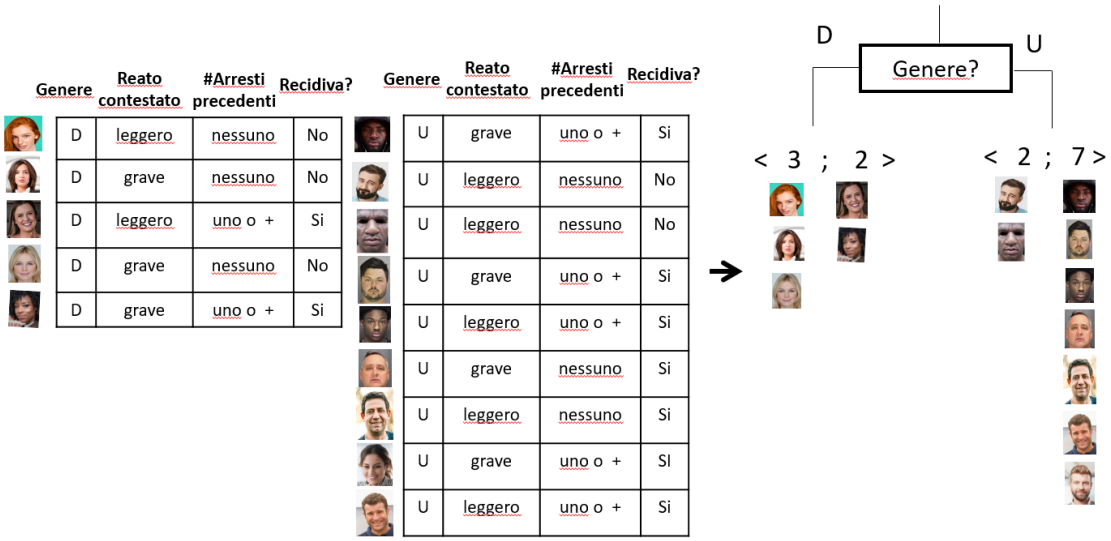


Figura 4.26 – Albero di decisione per la caratteristica Genere

Se uso la caratteristica Genere per separare gli elementi a basso rischio e ad alto rischio, ottengo come risultato i valori della Figura 4.27.

	# elementi della foglia	# elementi a basso rischio	# elementi ad alto rischio
Donne	5	3	2
Uomini	9	2	7

Figura 4.27 – Gli elementi a basso e alto rischio dell'albero di decisione di Figura 4.26.

A questo punto costruiamo gli alberi di decisione per ciascuna delle tre caratteristiche, vedi Figura 4.28, e una tabella che estende la tabella di Figura 4.27 a tutte e tre le caratteristiche, vedi Figura 4.29.

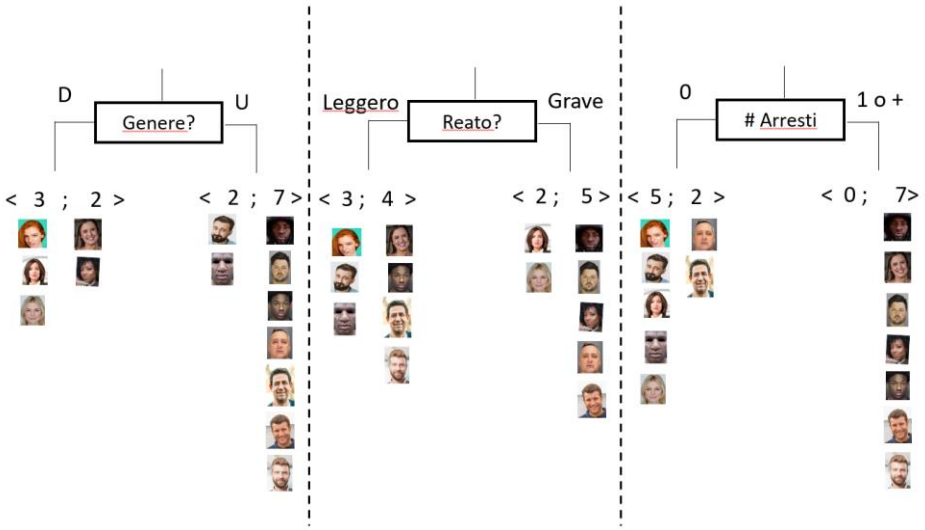


Figura 4.28 – I tre alberi di decisione generati dalle tre caratteristiche

	# elementi della foglia	# elementi a basso rischio	# elementi ad alto rischio
Donne	5	3	2
Uomini	9	2	7
Reato leggero	7	3	4
Reato grave	7	2	5
0 arresti	7	5	2
1 o + arresti	7	0	7

Figura 4.29 – Gli elementi a basso e alto rischio dell'albero di decisione per i tre alberi di decisione

A questo punto ti faccio una domanda: secondo te quale dei tre alberi di decisione separa di più gli elementi a basso e alto rischio? Se tu avessi a disposizione *una sola caratteristica* per separare in maniera più netta gli elementi a basso e ad alto rischio, quale sceglieresti?

Fammici pensare un attimo, ti do la risposta nella prossima pagina…

Non ho dubbi, sceglierei il numero di arresti: gli elementi con uno o + arresti sono tutti ad alto rischio!

Hai ragione!

Ho capito tutto. Esiste una formula matematica che esprima questo concetto di separazione?

Sì, e corrisponde al concetto di *entropia*, concetto che vuole proprio esprimere come siano separati gli elementi di una popolazione rispetto ai valori di una caratteristica.

Consideriamo come riferimento l'esempio di Figura 4.23 e seguenti. In generale l'entropia di una popolazione di n elementi (in Figura 4.23 n = 14) a cui corrispondono due valori (assumeremo: basso rischio; alto rischio), e che ha n1 elementi a basso rischio (nella formula usiamo il simbolo # ad indicare il numero degli elementi) e n2 elementi ad alto rischio è pari a

- [(#basso rischio/n) * logaritmo $_2$ (#basso rischio/n) + (#basso rischio/n) * logaritmo $_2$ (#basso rischio/n)]

Se l'insieme di Figura 4.23, composto di 14 elementi, avesse sette elementi a basso rischio e sette ad alto rischio l'entropia sarebbe

$$= - [(7/14) * \text{logaritmo}_2 (7/14) + (7/14) * \text{logaritmo}_2 (7/14)]$$

$$= - [(0.5) * \text{logaritmo}_2 (0.5) + (0.5) * \text{logaritmo}_2 (0.5)] =$$

$$- [0.5 * (-1) + 0.5 * (-1)] = 1$$

L'entropia = 1 è l'entropia massima che un insieme può avere e indica una situazione in cui la separazione tra i due gruppi è minima, ovvero, come si usa dire, il *disordine* è massimo, nell'esempio gli elementi sono sette di un tipo e sette dell'altro tipo.

Calcoliamo ora l'entropia dell'insieme di Figura 4.24:

$$= - [(9/14) * \text{logaritmo}_2 (9/14) + (5/14) * \text{logaritmo}_2 (5/14)] = 0.94$$

che è inferiore a 1 ma non di molto, perché 5 e 9 sono molto vicini a sette e sette.

Esercizio 4.5 - Calcola ora l'entropia (aiutati con un foglio excel…) per le sei foglie della Figura 4.28. Attenzione! Per calcolare l'entropia totale relativa a una certa caratteristica, ad esempio *genere*, l'entropia totale per donne e uomini, devi sommare le due entropie per donne e per uomini, pesandole con il numero di elementi; la formula è

Entropia associata a una caratteristica (ad esempio: Genere) = n1/n * entropia del nodo sinistro + n2/n * entropia del nodo destro

In pratica devi completare la tabella di Figura 4.30.

Valore della caratte-ristica	# ele-menti della foglia	# ele-menti a basso rischio	# ele-menti ad alto rischio	Entropia Foglia sinistra	Valore della caratte-ristica	# ele-menti della foglia	#ele-menti a basso rischio	#ele-menti ad alto rischio	Entropia Foglia destra	Entropia della caratte-ristica
Donne	5	3	2		Uomini	8	2	7		
Reato leggero	7	3	4		Reato grave	7	2	5		
0 arresti	7	5	2		1 o + arresti	7	0	7		

Figura 4.30 – Elementi quantitativi di partenza per il calcolo delle entropie

Vedi il risultato nella pagina successiva.

Soluzione dell'Esercizio 4.5

Dovresti aver ottenuto i valori della tabella in Figura 4.31.

Valore della caratte-ristica	# elementi della foglia	# elementi a basso rischio	# elementi ad alto rischio	Entropia Foglia sinistra	Valore della caratte-ristica	# elementi della foglia	# elementi a basso rischio	# elementi ad alto rischio	Entropia Foglia destra	Entropia della caratte-ristica
Donne	5	3	2	0.97	Uomini	8	2	7	0.76	0.83
Reato leggero	7	3	4	0.98	Reato grave	7	2	5	0.86	0.82
0 arresti	7	5	2	0.86	1 o + arresti	7	0	7	0	0.43

Figura 4.31 – Valori di entropia per le foglie e per le caratteristiche

L'entropia più bassa, 0.43, che corrisponde alla maggiore separazione tra basso rischio e alto rischio, è quella associata al *numero di arresti*.

La differenza di entropia tra la situazione iniziale (0.94) e l'albero di decisione associato al numero di arresti (0.43) è chiamata il *guadagno informativo*.

Mi puoi dare una definizione intuitiva di guadagno informativo?

Certo. Il guadagno informativo associato a un albero di decisione ad n livelli è il maggior grado di suddivisione, nel nostro caso tra elementi a basso rischio e elementi ad alto rischio, che l'albero di decisione manifesta nell'insieme delle sue foglie, rispetto all'albero considerato in precedenza ad n – 1 livelli.

Possiamo ora fornire un metodo che permette di costruire un albero di decisione efficiente nel separare gli elementi a basso rischio rispetto a quelli ad alto rischio, nel caso, usuale in cui abbiamo a disposizione tante caratteristiche, vedi Figura 4.32.

Dobbiamo:

- Primo passo - calcolare i guadagni informativi per tutte le caratteristiche
- Secondo passo - scegliere come prima caratteristica dell'albero di decisione quella con maggior guadagno informativo.
- Terzo passo - a questo punto ricalcoliamo i guadagni informativi e procediamo nella scelta delle ulteriori caratteristiche
- Quarto e ultimo passo – fino a quando non possiamo più ottenere guadagni informativi significativi.

	Genere	Reato contestato	#Arresti precedenti	C4	C5	C6	C7	C8	C9	C10	C11	Recidiva?
 1	D	leggero	nessuno	..	..	..	..	..	..	..	..	No
 2	U	grave	uno o +	..	..	..	..	..	..	..	..	Si
 3	U	leggero	nessuno	..	..	..	..	..	..	..	..	No
 4	D	grave	nessuno	..	..	..	..	..	..	..	..	No
 5	D	leggero	uno o +	..	..	..	..	..	..	..	..	Si
 6	U	leggero	nessuno	..	..	..	..	..	..	..	..	No
 7	U	grave	uno o +	..	..	..	..	..	..	..	..	Si
 8	D	grave	nessuno	..	..	..	..	..	..	..	..	No
 9	D	grave	uno o +	..	..	..	..	..	..	..	..	Si
 10	U	leggero	uno o +	..	..	..	..	..	..	..	..	Si
 11	U	grave	nessuno	..	..	..	..	..	..	..	..	Si
 12	U	leggero	nessuno	..	..	..	..	..	..	..	..	Si
 13	U	grave	uno o +	..	..	..	..	..	..	..	..	SI
 14	U	leggero	uno o +	..	..	..	..	..	..	..	..	Si

Figura 4.32 – Una popolazione di cui sono note tante caratteristiche

4.3 Prime conclusioni sulle tecniche di Machine learning partendo da esempi

Siamo arrivati a un buon punto per trarre alcune conclusioni. Se ti è tutto chiaro, puoi saltare questa sezione. Altrimenti vedila come un ripasso.

In questo libro siamo interessati a approfondire i risvolti etici connessi all'uso di *modelli predittivi* che ci permettono di prevedere un fenomeno nel futuro, nel nostro esempio il rischio di commettere recidiva. Abbiamo capito questo:

1. Il *Machine learning* partendo da esempi è una *disciplina nell'ambito della Intelligenza Artificiale* che permette di utilizzare
2. una *tecnica di apprendimento*, come per esempio gli alberi di decisione, per costruire
3. un *modello predittivo* che ci consente di prevedere il livello di rischio con cui
4. ad esempio, un *detenuto* che ottiene la libertà provvisoria commetterà o meno in futuro recidiva.

Tu hai sottolineato quattro termini: Machine learning, tecnica di apprendimento, modello predittivo e ….detenuto. Mi sembra che detenuto non c'entri molto con gli altri. Perché?
Ottima domanda: *c'entra, eccome!!!* Tutta la strumentazione di tecniche connesse al Machine learning produce modelli classificatori e predittivi che riguardano le *persone*, questo è il punto fondamentale del libro: quale etica avere come riferimento per tecniche di apprendimento e modelli classificatori e predittivi che riguardano le persone?
Ci arriveremo nell'ultimo capitolo….

Quando costruiamo un modello predittivo, noi assumiamo di conoscere alcune *caratteristiche descrittive* degli elementi del campione, come ad esempio il genere, e una *caratteristica obiettivo,* nel nostro caso il rischio di recidiva. Assumeremo due valori possibili per la caratteristica obiettivo, rischio basso e rischio alto.

Esistono varie tecniche di apprendimento che permettono di costruire modelli predittivi, ad esempio gli alberi di decisione.

Un *albero di decisione* è costruito a partire da un campione di elementi (ad esempio un insieme di detenuti) per fasi successive, creando nodi e diramazioni dai nodi, utilizzando una caratteristica per volta (ad esempio prima il genere e poi il reato contestato)[4] e associando gli elementi del campione ai vari rami a seconda dei valori delle caratteristiche (ad esempio, le donne associate al ramo a sinistra e gli uomini associati al ramo a destra); tutto ciò fino ad arrivare, esaurite le caratteristiche considerate, alle foglie dell'albero, cui si associano il numero di elementi della foglia con rischio basso e rischio alto.

Sulla base dei due valori associati alle foglie (numero di elementi con rischio basso e numero con rischio alto), si sceglie per ogni foglia il valore più alto e si associa la previsione a tale valore più alto. Ad esempio:

- *se* alla foglia associata alle donne accusate di reato leggero, con uno o più arresti precedenti, sono associati tre elementi che non hanno commesso recidiva e un elemento che l'ha commessa <3;1>, si associa alla foglia *rischio basso*
- per ogni nuovo detenuto che fa domanda di libertà provvisoria*,* e le cui caratteristiche corrispondono alla foglia, si associa al detenuto *rischio basso, e* conseguentemente si concede al detenuto la libertà provvisoria.

Si capisce? Ti torna?

Sì, si capisce, il ripasso è stato utile, ho capito anche meglio gli esempi del Capitolo 1. A questo punto però ho alcune domande da fare. Per cercare di essere chiaro, darò un titolo a ogni domanda.

<u>Prima domanda: secondo me, la procedura può portare a commettere degli errori o delle ingiustizie.</u> Tu l'hai già fatto notare, ma ci voglio tornare. Facciamo una specie di gioco, e applichiamo l'albero di decisione in Figura 4.9 ai quattordici casi di Figura 4.6, e in particolare al tredicesimo elemento, quello che associa a <uomo, reato grave e uno o + arresti pregressi> il non aver commesso recidiva. Ebbene, se sottoponiamo quel caso all'albero di decisione di Figura 4.12, otteniamo un valore di alto rischio, anche se l'elemento non ha commesso recidiva, e ciò perché prevalgono gli altri due che avevano commesso recidiva, insomma: due pesi e due misure! La stessa cosa succede quando ci si trova nella situazione di Figura 4.17, in cui abbiamo il caso <1;1>: se decidiamo basso rischio, favoriamo il "secondo 1", se decidiamo alto rischio, sfavoriamo "il primo 1". Come la mettiamo?

Eh, la mettiamo che hai ragione! Hai constatato che gli alberi di decisione possono portare a decisioni sbagliate o non eque. E siccome questo è il tema dell'intero libro, la equità, ti chiedo di pazientare un po' per una risposta completa alla tua osservazione. Ti propongo ora un altro esercizio.

[4] Esistono anche alberi di decisione in cui possiamo considerare più di una caratteristica per volta.

Esercizio 4.6 - Per cominciare a ragionare sulla questione degli errori ti dò un esercizio. Guarda la Figura 4.33, in cui si suppone di considerare una sola caratteristica, il numero di arresti precedenti, e due possibili ramificazioni della decisione, come mostrato in figura.

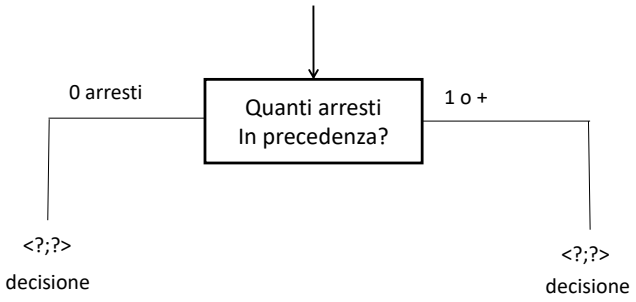


Figura 4.33 – Albero di decisione in cui si considera una sola caratteristica e due nodi foglia

Dovresti calcolare:
1. le frequenze dei valori rischio basso e rischio alto per i due nodi foglia nel caso dell'insieme dei quattordici elementi di Figura 4.23.
2. le relative decisioni per i due nodi tra rischio alto e rischio basso, e quindi su no-libertà provvisoria e si-libertà provvisoria
3. gli errori che si commettono qualora usiamo l'albero di decisione così costruito sui 14 elementi.

Le risposte nella prossima pagina.

Soluzione all'Esercizio 4.6 - L'albero di decisione con frequenze e decisioni è mostrato in Figura 4.34.

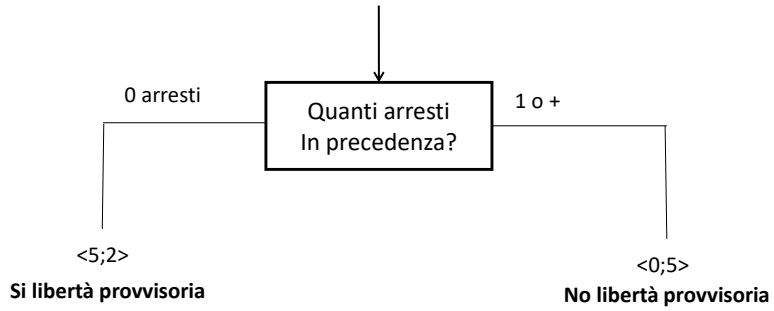


Figura 4.34 - Albero di decisione soluzione dell'esercizio

Gli errori che si commettono sono:
nel caso di 0 arresti: 2 errori su 7
nel caso di uno o + arresti: nessun errore.

Bene, per ora ho capito.

Seconda domanda: negli alberi di decisione come tecnica per il Machine learning da esempi, io non vedo nessun apprendimento; hai scelto delle caratteristiche, hai selezionato degli esempi dal passato, hai fatto un po' di statistiche sugli esempi, e poi hai costruito un albero di decisione, questo è tutto! Non c'è nessun apprendimento! Come la mettiamo?

Altra domanda centrata! Dunque, anzitutto per far comprendere la logica della più semplice delle tecniche di Machine learning, gli alberi di decisione, ho usato un approccio semplificato, quello che tu hai ben descritto. Per rispondere alla tua domanda occorre dire che:

1. come si è visto dagli esempi, il processo di selezione delle caratteristiche e dell'ordine con cui considerarle basato sul guadagno informativo porta a una complessa iterazione di passi ed esperimenti che generano alberi sempre più precisi nell'identificare i casi possibili e quindi sempre più efficaci nel discriminare le previsioni sulla caratteristica obiettivo; questi passi sono certamente una forma di apprendimento.
2. Io ho potuto fare casi molto semplici, costituiti da al massimo ventidue elementi in tutto. Nella realtà vengono usati insiemi di dati costituiti da un numero di elementi che può arrivare all'ordine dei milioni e anche dei miliardi, con centinaia di caratteristiche. Si può intuire che al crescere del numero degli elementi l'albero di decisione porterà a decisioni meno "locali" e più "generali", insomma non avremo più gli <1;0>, <1;1> ecc. che abbiamo visto nei casi che ho mostrato, ma statistiche meno contingenti e valide per un gran numero di elementi; si può vedere che al crescere del numero degli elementi la precisione della decisione cresce sempre più. Anche questo è apprendimento.
3. Io ho mostrato la più semplice delle tecniche per Machine learning, gli alberi di decisione, ce ne sono tantissime altre, e, anche qui mi affido alla tua intuizione, ce ne sono molte in cui l'apprendimento diventa una caratteristica costitutiva della tecnica. Per esempio, si possono usare contemporaneamente, come nel Machine learning ensemble, tante tecniche, nel nostro caso, tanti alberi di decisione invece che uno solo, e decidere la

tecnica ottima mediante una votazione, in cui si sfrutta il meglio dei vari alberi. Nel caso della libertà provvisoria, si può scoprire che per alcuni detenuti ha maggiore capacità predittiva la tipologia dei reati che hanno portato ad arresti precedenti, mentre per altri può avere maggior peso l'età del detenuto, e così via. Vedremo una di queste tecniche, le foreste casuali. nel prossimo capitolo.

Conclusioni sul Machine learning partendo da esempi

Abbiamo fatto tanta strada in questo capitolo!

Si, tanta strada….

Le diverse fasi della produzione di modelli predittivi sono mostrate in Figura 4.35, in cui ho distinto tre passi, al solito esemplificati per il caso Compas.

Il primo passo *osserva il passato*; viene raccolta tutta la conoscenza disponibile sul comportamento di un insieme di detenuti che nel passato hanno ottenuto la libertà provvisoria.

In questa fase, partendo dalle informazioni disponibili su quanto accaduto nel passato, *scegliamo le caratteristiche* da prendere in considerazione per costruire il modello predittivo Per esempio, è possibile che nel passato siano state raccolte informazioni sull'ultimo titolo di studio raggiunto e sul colore degli occhi del detenuto. Ora, mentre il titolo di studio è un elemento che può aver influenzato il comportamento della persona, non si capisce come il colore degli occhi possa influire. Quindi, possiamo decidere di tenere in considerazione la prima caratteristica e trascurare la seconda.

Nel secondo passo:
- scegliamo la *tecnica di apprendimento* e
- costruiamo, a partire dai dati disponibili del passato, il *modello predittivo*.

Il modello consiste in un albero di decisione in cui sono stati calcolati i valori della caratteristica obiettivo per tutti i casi che corrispondono alle foglie dell'albero.

I risultati del modello predittivo conseguenti sono mostrati in Figura 4.35 per mezzo di una tabella, in cui le ultime due colonne rappresentano il livello di rischio e la decisione.

A questo punto, terzo passo, utilizziamo il modello predittivo per prevedere il livello di rischio di recidiva per i detenuti attuali, e per prendere la decisione definitiva sulla concessione o meno della libertà provvisoria. Nel quarto passo si prende la decisione sui detenuti.

Coloro che guardano con perplessità l'uso del Machine learning per aiutare gli esseri umani a formulare previsioni affermano che i modelli e le tecniche non tengono conto di tanti aspetti che *non sono definibili per mezzo di dati*. E su questo mi sento assolutamente di dare loro ragione.

Tuttavia, mentre i decisori umani hanno un limite cognitivo nell'elaborare una grande quantità di conoscenza, le tecniche di Machine learning sono in grado di elaborare rapidamente immense quantità di dati. Il futuro della ricerca dovrà farci capire come umani e tecniche possono collaborare insieme raggiungendo il massimo di efficacia.

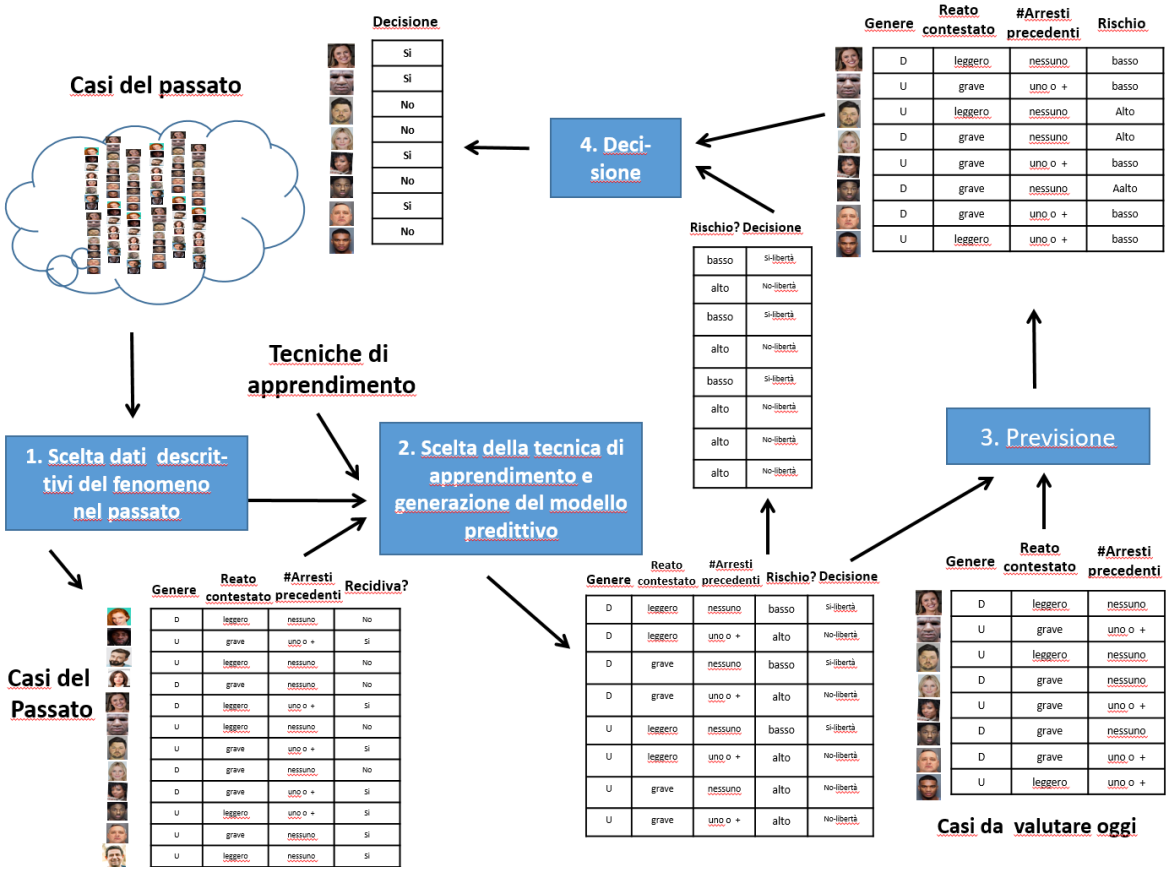


Figura 4.35 – Il ciclo di vita del Machine learning partendo da esempi

Capitolo 5 - L'equità nella filosofia, nelle scienze giuridiche, nella vita sociale

"Perché ha fatto passare davanti a me quella signora? Perché era più anziana, e era visibilmente affaticata….E perché ha fatto passare avanti anche quel giovane? Eh, è un mio amico…."
"Il professore mi ha dato 27, ma meritavo di più. A me ha fatto domande molto difficili, allo studente prima ha fatto domande molto più facili, e gli ha dato 30, non è giusto!"
"Quel giudice ha dato una pena eccessiva all'imputato"

I casi precedenti sono esempi di decisioni formulate da esseri umani, non legate all'utilizzo di modelli predittivi, in cui noi percepiamo la presenza di discriminazioni. Le decisioni riguardano chi servire in un negozio, il voto a un esame universitario, la condanna e la pena conseguente o la assoluzione in una sentenza penale; ogni decisione, soprattutto quelle importanti, accontentano qualcuno e scontentano altri.

Il concetto di equità nelle previsioni e decisioni sembra semplice da formulare: una previsione o decisione è equa se è giusta, se non favorisce nessuno. Ma appena cerchiamo di comprendere il concetto di giustizia, subito ci vengono in mente storie e contesti che minano fortemente la nostra capacità di dare definizioni oggettive.

5.1. Le lotte per la non discriminazione degli Afro-Americani negli Stati Uniti

Rosa Parks[5] era cittadina di Montgomery ed era *Afro-Americana*. Così era stata definita la nuova razza nata dall'importazione di schiavi nella "terra della libertà" sin dalla sua nascita. Rosa Parks, al ritorno dal lavoro (era rammendatrice presso una sartoria) il 1 Dicembre 1955

Figura 5.1 - Discriminazione razziale neglòi Stati Uniti

[5] Il testo che segue su Rosa Parks è tratto da Wikipedia

salì sull'autobus che era particolarmente affollato, vista l'ora. L'unico posto libero era nella parte centrale, di fatto accessibile alla popolazione sia bianca che nera da una specifica legge in vigore nella cittadina di Montgomery.

I posti a sedere negli autobus erano di tre categorie diverse: i primi dieci posti nella parte anteriore erano riservati ai *Bianchi*, i dieci posti nella parte posteriore erano riservati ai *neri* e i sedici posti nella parte centrale dell'automezzo erano di uso misto, nel senso che persone di entrambe le razze potevano utilizzarli ma, se eventualmente i posti non fossero stati sufficienti per tutti, un nero aveva il dovere di lasciare il posto libero al cittadino di razza bianca.

Allorché un cittadino bianco salì sull'autobus, non trovando posto, il conducente del mezzo ingiunse alla donna di alzarsi. Rosa oppose rifiuto e venne richiesto così l'intervento della polizia che procedette al suo arresto con l'accusa di aver violato le leggi sulla segregazione.

La legge sui diritti civili degli Stati Uniti del 1964 mise al bando le discriminazioni sulla base della razza, il colore, la religione, il sesso o origine nazionale di un individuo. La legge conteneva due importanti disposizioni che esprimevano la comprensione della comunità dei cittadini su cosa significava essere non equi: il titolo VI, che impediva alle agenzie governative (comprese le università) di ricevere fondi federali che discriminavano in base alla razza, colore o origine nazionale; e il titolo VII, che impediva ai datori di lavoro con 15 o più dipendenti di discriminare nei rapporti di lavoro in base alla razza, colore, religione, sesso o origine nazionale.

Partiremo dunque dal gesto di Rosa Parks per indagare il concetto di equità nella filosofia, nelle scienze giuridiche, nelle legislazioni.

5.2 L'equità e il suo opposto, la discriminazione, nella filosofia e nelle scienze giuridiche

L'equità è oggetto di studio dalla nascita della filosofia e dei codici giuridici che hanno regolato le controversie tra gruppi sociali e tra le singole persone. Vediamo senza pretesa di completezza come sia considerata nei due ambiti.

Equità e obiettività in filosofia

Un' ottima introduzione al concetto di obiettività (o oggettività) in filosofia la troviamo nel libro[6], pubblicato in uno di quegli autentici capolavori di sintesi e chiarezza che sono i libri della serie "Very short introductions" della Oxford University Press. Manterrò il termine *obiettività* discusso nel testo, ma risulterà chiaro da quanto segue che esso è molto vicino al concetto di equità investigato in questo libro.

Tre sono le possibili accezioni del termine obiettività. La prima vede la obiettività come *assenza di pregiudizi e di distorsioni nella nostra interpretazione del mondo*. Una previsione o decisione obiettiva è una valutazione che può essere condivisa da ogni persona imparziale,

[6] Stefen Gaukrogher – Objectivity, A very Short Introduction, Oxford University Press, 2012

indipendentemente dalle diverse visioni del mondo che le persone possono avere. Questa accezione colloca la obiettività in un contesto molto concreto, il *contesto della società*.

Ciò è in contrasto rispetto a come è intesa la obiettività nel dominio della scienza. L'idea della scienza, caratterizzata dal rigore che chiede agli scienziati nel verificare sperimentalmente i risultati delle loro attività di ricerca, non è incorporata in questa idea di obiettività. Non è infatti realistica, se non in frange molto limitate, la visione dello scienziato come soggetto che si deve liberare dai pregiudizi per raggiungere la obiettività nella sua attività di ricerca.

Questa prima accezione della obiettività sarà la accezione considerata in prevalenza in questo libro.

La seconda accezione di obiettività vede una previsione o decisione come obiettiva quando è *libera da ipotesi fatte a priori, o sistemi di valori* che la influenzano. Apparentemente questa accezione è una estensione della precedente. Si potrebbe infatti dire: chi può affermare che le visioni del mondo, le ipotesi che formuliamo sul mondo, non siano pregiudizi e distorsioni?

La risposta a questa domanda consiste nell'osservare che l'idea intuitiva che sta dietro al concetto di pregiudizio porta a vederlo come una distorsione del nostro giudizio sul mondo, distorsione che ci offusca la capacità di discernimento; mentre l'idea che sta dietro ai concetti di ipotesi sul mondo, di sistemi di valori attraverso cui osserviamo il mondo, non lo è. Una cosa sono dunque le distorsioni, che razionalmente possiamo rimuovere, una cosa sono i valori, che sentiamo più nostri, e da cui, certamente, è molto più difficile distaccarsi.

In effetti:
- nel primo caso, *assenza di distorsioni*, gli argomenti che uno utilizza in un ragionamento, e che *non sono condivisi*, dovrebbero essere rimossi se il ragionamento e le sue conclusioni sono da considerarsi obiettivi.
- nel secondo caso, *assenza di assunzioni a priori*, gli argomenti che uno utilizza in un ragionamento, *sia che siano condivisi sia che non lo siano*, dovrebbero essere *rimossi*.

La differenza tra le due accezioni *è fondamentale*: molti degli argomenti degli scettici e dei relativisti fondono e non distinguono tra la prima e la seconda accezione, così che il compito di rimuovere tutti i *pregiudizi*, *raggiungibile* con una severa autoanalisi, è trattato alla stregua del compito di rimuovere tutte le *assunzioni a priori*, compito questo molto più arduo e talvolta *irraggiungibile*.

La terza accezione di obiettività considera una previsione o decisione obiettiva quando ci permette di *decidere tra visioni del mondo o teorie in conflitto tra di loro*. Mentre le prime due accezioni fanno riferimento a uno stato della mente, essere liberi da pregiudizi, o da assunzioni, questa terza accezione fa qualcosa di differente. Definisce un percorso, una procedura, che noi dobbiamo seguire per decidere tra punti di vista differenti. E' questa la accezione *adottata tipicamente nella scienza*.

Karl Popper, che vedeva la obiettività e la scienza come concetti equiparati, collocati allo stesso livello, affermava che nella scienza le nostre ipotesi hanno conseguenze che forniscono evidenze empiriche, per cui noi abbiamo sempre la possibilità di verificarle su ciò che accade

nella realtà. Nella storia, al contrario, i fatti sono sempre dietro di noi, nel passato, per cui non possiamo mai confrontare le nostre evidenze con i fatti.

Il punto rilevante qui è che l'obiettività ci richiede di concepire procedure per poter decidere tra valutazioni in competizione, allo scopo di scoprire quale porti a predizioni corrette, e ci impone di scegliere quella che effettua predizioni *fattualmente* corrette, cioè verificate con i fatti.

Come le predizioni siano formulate e verificate è un argomento centrale in questo libro. Affronteremo molte questioni sui dati a partire dai quali formulare le previsioni, sui modelli che ci permettono di formulare previsioni partendo dai fatti del passato, e vedremo quanto sarà complesso questo cammino anche nella scienza del Machine learning, quella che consideriamo in prevalenza in questo libro.

Ciò che viene sottolineato nel testo che ho scelto come riferimento è che in questa terza asserzione del concetto di obiettività ci troviamo di fronte piuttosto a una *condizione necessaria* per l'obiettività, non a una definizione della obiettività *in sé*. Questo criterio, insomma, non ci dice cosa sia la obiettività, come cercano di fare le prime due asserzioni, quanto piuttosto ci propone *ciò che dobbiamo fare* per raggiungere l'obiettività, consapevoli che può non bastare, che può essere solo una condizione necessaria, ma non sufficiente.

Quando vedremo le diverse forme di equità su cui può essere valutato un modello predittivo, vedremo che, effettivamente, non ci troveremo di fronte a un solo tipo di *equità* (usiamo qui il termine investigato nel libro), ma tanti tipi, in contrasto tra loro, e vedremo come questa terza forma di obiettività non sia solo una astratta speculazione filosofica, ma uno strumento fondamentale per affrontare quella che a buon diritto possiamo chiamare la *confusione del mondo*.

L'equità e la discriminazione nelle scienze giuridiche: una breve introduzione

Il concetto di equità, con riferimento al suo contrario, la discriminazione, viene discusso secondo una prospettiva giuridica nel testo[7], cui farò riferimento nel seguito.

Chiamare qualcosa *discriminazione*, in molti contesti, significa asserire che una determinata decisione porta ingiustamente un *danno* a una persona, *e non ad altre*, che pure dovrebbero avere lo stesso trattamento, ovvero un *beneficio* a una persona, *e non ad altre*, che pure ne avrebbero diritto.

In alternativa, a volte il termine discriminazione ha un significato meno negativo, per cui la azione cui fa riferimento non è necessariamente dannosa o sbagliata. I requisiti per cui si devono avere 16 anni per ottenere la patente per motocicli e 18 per votare sono esempi di discriminazione basata sull'età, che però tutti accettiamo, in quanto ispirati a un principio di prudenza, nel caso della guida, o maturità nel caso del diritto di voto.

[7] Deborah Hellman - When Is Discrimination Wrong? – Harvard University Press, 2008.

Nel seguito siamo interessati al primo significato del termine. D'ora in poi userò il termine *discriminazione* o il suo opposto, *equità*, a seconda delle circostanze, quando intendo sottolinere gli aspetti negativi o positivi del concetto. Utilizzerò anche i termini *iniquità* e *distorsione* come sinonimi del termine discriminazione.

Esistono almeno due tipi di equità. L'equità *comparativa* fonda il giusto trattamento di una persona, o di un gruppo di persone, sulla base di un confronto con altre persone, o altri gruppi di persone. Ad esempio, "le donne guadagnano meno degli uomini" è un esempio di iniquità comparativa. La equità *non comparativa* fonda il giusto trattamento di una persona, o di un gruppo di persone, senza considerare come vengono trattate le altre persone o gruppi di persone, ma solo sulla base di un criterio universale di riferimento.

Per esempio, supponiamo che un professore universitario si dia come regola per i voti degli esami di dare uniformemente in pari proporzioni agli studenti promossi tutti i voti da 18 a 30. In questo caso non rispetta la equità non comparativa, perché il voto non è dato sulla base del merito (potrebbe accadere che tutti gli studenti hanno risposto correttamente a tutte le domande, e meritano tutti 30); invece, nell'ambito della equità comparativa può applicare quella regola, ma assegnando lo stesso voto a compiti di uguale qualità.

Abbiamo visto che ci sono dunque varie forme di discriminazione/equità:
- la discriminazione imposta sulla base di leggi condivise, messa in atto non per sfavorire ma per proteggere una fascia sociale di popolazione: ad esempio, non possono avere la patente per i motocicli i giovani con meno di 16 anni.
- la discriminazione/equità *etica*, che consiste nel diverso/uguale trattamento di singole persone o gruppi di persone, per esempio donne o uomini, che a sua volta può essere:
  - *comparativa,* quando dà luogo a un *diverso/pari* trattamento di due persone o gruppi di persone.
  - *non comparativa*, quando dà luogo a un trattamento *iniquo/equo* di una persona o un gruppo di persone rispetto a un criterio di riferimento.

La distinzione tra discriminazione/equità comparativa e non comparativa esiste anche nel mondo degli studiosi e dei magistrati che operano nell'ordinamento giudiziario. Per *ordinamento giudiziario*, in diritto, si intende l'insieme delle norme che regolano l'attività giurisdizionale, cioè l'attività di risoluzione delle controversie da parte di giudici terzi e imparziali, che al fine di risolvere le controversie a loro presentate interpretano e applicano le leggi.

Alcuni studiosi del diritto ritengono che debba essere la giustizia non comparativa a governare il processo di generazione delle sentenze. Una persona giudicata colpevole, dovrebbe avere la punizione che merita sulla base del reato che ha compiuto e delle circostanze attenuanti o aggravanti, non con riferimento ad altri.

Altri studiosi del diritto ritengono che debba essere applicata la giustizia comparativa, che si basa ad esempio sui casi *precedenti* simili. Questo principio, come abbiamo già visto, ispira tutti i modelli predittivi basati su apprendimento. Noi nel seguito saremo dunque interessati alla discriminazione/equità comparativa nell'ambito dei modelli predittivi generati da tecniche di Machine learning.

Ma la questione della equità e della discriminazione non è così semplice. Cosa succede quando una ipotetica norma stabilisse che le persone che vogliono prendere la patente devono pagare 500 euro? E' questo un esempio di discriminazione comparativa? Apparentemente no, perché quella norma tratta tutti i cittadini allo stesso modo; in realtà, se è vero che c'è uguale trattamento verso tutti i cittadini (poteva accadere che le persone con gli occhi azzurri dovessero pagare 100 euro e quelle con occhi neri 500 euro…) l'impatto sui portafogli dei cittadini è diverso per le persone indigenti e per le persone benestanti.

Esistono dunque norme che ci appaiono non discriminatorie (patente per motociclo solo se sei maggiore di 18 anni), e norme che invece ci appaiono discriminatorie/inique (pagare 500 euro per la patente). Questo significa che oltre al trattamento, noi dobbiamo anche guardare all'impatto sulle singole persone: se, nel caso precedente, il trattamento è equo, l'impatto è discriminatorio.

Possiamo dunque dire che una decisione, una norma, un uso discriminano in modo diretto una persona A (o un gruppo di persone A) se essa è trattata differentemente da un'altra persona B (o gruppo di persone B) sulla base del fatto che essa possiede o non possiede una certa caratteristica (es. è uomo o donna). Si parla in questo caso di *diverso trattamento*.

Una decisione, una norma, un uso discriminano in modo indiretto una persona (o un gruppo di persone) se essa ha un impatto diverso su persone che possiedono o non possiedono una certa caratteristica. Si parla in questo caso di *diverso impatto*.

Allo stesso tempo, è evidente che il confine nelle norme tra quelle che ci appaiono discriminatorie e non discriminatorie è molto labile, cosa diremmo per un pagamento di 200 euro, di 50 euro, di 10 euro ecc.?

E anche tra le norme che abbiamo associato a leggi condivise, se in un paese esiste una legge che stabilisce che hanno diritto di voto solo gli uomini e non le donne, c'è poco da discutere, con la nostra sensibilità sociale noi consideriamo la legge gravemente discriminatoria verso le donne. E consideriamo l'organo che detiene il potere legislativo, il Parlamento di quel paese, antidemocratico (se non altro perché è stato eletto solo da uomini…).

Infine, occorre ricordare che spesso le discriminazioni non scaturiscono soltanto da leggi, ma anche da norme organizzative, ordini di servizio, regole sociali consolidate, usi, credenze. Che succede se un amministratore delegato di una azienda decide di assumere solo uomini, perché le donne possono assentarsi a causa di una gravidanza, e quindi fanno diminuire i profitti della azienda?

Le lavoratrici, i lavoratori in generale possono considerare la precedente come una decisione discriminatoria, e se non c'è una legge che li protegge, possono decidere di contrastare questa discriminazione con lo strumento dei contratti collettivi o contratti aziendali. *Insomma, la discriminazione/equità è una tema che permea tutta la nostra vita, e che spesso è divisivo rispetto alle nostre opinioni sui diritti sociali e alle nostre opinioni politiche.*

Ciò che è certo è che in genere noi associamo una discriminazione a una norma o a una decisione, sulla base del fatto che le persone investite dalla norma o dalla decisione posseggano o meno una certa caratteristica. Ma anche questa affermazione va approfondita.

Per esempio, confrontiamo:

- il Caso 1, in cui uno studio di avvocati assume Pietro e non Giovanni, perchè Pietro è bianco e Giovanni è Afro-americano, con il
- Caso 2, in cui l'azienda assume Pietro e non Giovanni sulla base dell'esito di un esame in cui Pietro ha preso 100 e Giovanni 50, e infine il
- Caso 3, in cui l'azienda assume Pietro e non Giovanni sulla base dell'esito di un modello predittivo automatico basato su Machine learning, che ha confrontato i curricula di Pietro e Giovanni per capire se abbiano caratteristiche simili agli impiegati assunti nel passato ovvero quelli non assunti.

Come si spiega il convincimento etico che nel Caso 1 ci spinge a considerare la decisione discriminatoria e nel Caso 2 non discriminatoria? La risposta alla domanda sembra semplice.

- Si può ritenere il Caso 1 discriminatorio perché la etnia non è rilevante nella professione di avvocato, mentre la votazione all'esame lo è.
- Si può presumere che nel Caso 1 l'esaminatore sia stato mosso da un sentimento ostile verso l'Afro-Americano ovvero da una credenza radicata che gli Afro-Americani siano meno bravi, e ciò rende la scelta indifendibile.

Mi fermo qui nel trattare il tema della discriminazione/equità dal punto di vista che ho chiamato delle scienze giuridiche; il lettore interessato ad approfondire questi temi può leggere il bellissimo libro segnalato in precedenza.

Scusa, ma ti sei scordato di commentare il Caso 3 di prima!

Certo, me lo sono scordato perché lo svilupperemo nel resto del libro! Già fin da adesso possiamo comprendere che nel caso di modelli basati su Machine learning per comprendere se discrimina o non discrimina dovremo capire come il modello "ragiona". Per ora possiamo arrivare alle seguenti conclusioni.

Le discriminazioni/equità possono avere una natura molto varia. Anzitutto ci sono forme di discriminazioni stabilite da leggi condivise, accanto ad esse discriminazioni etiche che dipendono da come, *nella decisione a cui la discriminazione è associata*, vengono prese in considerazione determinate caratteristiche dei soggetti cui si applica la decisione. Le discriminazioni possono riguardare singole persone o popolazioni di persone; possono generare un diverso trattamento o un diverso impatto, e infine possono essere di natura comparativa, o non comparativa.

L'equità e la discriminazione nelle legislazioni: introduzione

Occorre infine osservare che le società nel corso della loro storia sono arrivate alla conclusione che determinate forme di discriminazione sono a tal punto contrarie ai diritti dei cittadini, o a una loroparte, che *debbano essere disciplinate da leggi*, e non dal libero

convincimento di coloro che le attuano; e che, se coloro che le generano violano tali leggi anti discriminazione, debbano, qualora tale atteggiamento sia accertato da un procedimento amministrativo o giudiziario, pagare una pena.

Nelle prossime sezioni, esamineremo, nel caso degli Stati Uniti e della Unione Europea, quali siano stati i provvedimenti legislativi principali di contrasto alle diverse forme di discriminazione. Per gli Stati Uniti focalizzeremo l'attenzione sul diritto al lavoro.

5.3. Ambiti legislativi sul diritto al lavoro negli Stati Uniti

Le leggi in tema di uguale opportunità di impiego negli Stati Uniti rendono illegale per le agenzie federali discriminare i dipendenti e i candidati al lavoro sulla base di razza, colore, religione, sesso, origine nazionale, disabilità o età. Una persona che presenta un reclamo o partecipa a un'indagine su un reclamo, o che si oppone a una pratica lavorativa resa illegale ai sensi di una qualsiasi delle leggi in tema di uguale opportunità di impiego, è protetta da ritorsioni.

Il titolo VII del Civil Rights Act del 1964 protegge i dipendenti e le persone in cerca di lavoro dalla discriminazione sul lavoro basata su razza, colore, religione, sesso e origine nazionale. La protezione del titolo VII copre l'intero spettro delle decisioni in materia di assunzione, comprese le assunzioni, selezioni, licenziamenti e altre decisioni relative a termini e condizioni di lavoro. Ad esempio, la discriminazione razziale consiste nel trattamento sfavorevole di qualcuno (un candidato o dipendente) perché appartiene a una determinata razza/etmia o a causa di caratteristiche personali associate alla razza/etnia (come la capigliatura, il colore della pelle o alcuni tratti del viso). La discriminazione del colore comporta il trattamento sfavorevole di qualcuno a causa del colore della carnagione della pelle.

La discriminazione di razza/etnia o colore avviene anche quando vi sia trattamento sfavorevole di qualcuno perché la persona è sposata (o associata) a una persona di una certa razza o colore.

L'Equal Pay Act del 1963 protegge uomini e donne dalla discriminazione salariale basata sul genere nel pagamento di salari o benefici a persone che svolgono un lavoro sostanzialmente uguale nello stesso istituto.

L'Age Discrimination in Employment Act (ADEA) del 1967 protegge le persone di età pari o superiore a 40 anni dalla discriminazione sul lavoro basata sull'età.

Le sezioni 501 e 505 del Rehabilitation Act del 1973 proteggono i dipendenti e i candidati al lavoro dalla discriminazione sul lavoro basata sulla disabilità.

Il Civil Rights Act del 1991 modifica diverse sezioni del Titolo VII per rafforzare e migliorare le leggi federali sui diritti civili e prevedere il recupero dei danni compensativi nel settore federale per discriminazione intenzionale sul lavoro.

Altri ambiti coperti nella legislazione e nei regolamenti di specifiche agenzie negli Stati Uniti riguardano:

- Assegnazione di alloggi
- Concessione di prestiti bancari
- Concessione di benefici sociali
- Concorsi e assunzioni
- Carriere
- Promozioni a scuola
- Polizia predittiva
- Fasi del procedimento penale
- Sentenze di procedimenti giudiziari civili.

5.4 Non discriminazione nella legislazione della Unione Europea

Nella Unione Europea l'obiettivo della legislazione in materia di non discriminazione è quello di offrire a tutte le persone possibilità di accesso pari ed eque alle opportunità disponibili nell'ambito di una società. Ciò implica che le persone e i gruppi di persone non siano trattati in maniera meno favorevole in presenza di situazioni equiparabili solo a causa di caratteristiche particolari, tra cui genere, razza, origine etnica, religione o convinzioni personali, disabilità, età o orientamento sessuale.

Il trattato sul funzionamento dell'Unione europea (TFUE) vieta la discriminazione in base alla nazionalità. Esso consente inoltre al Consiglio europeo di adottare provvedimenti opportuni per lottare contro le discriminazioni fondate su sesso, razza o origine etnica, religione o convinzioni personali, disabilità, età o orientamento sessuale. A tale proposito, il Consiglio deve agire all'unanimità e previo l'ottenimento dell'approvazione del Parlamento europeo. Tuttavia, nell'ambito specifico della parità di trattamento e delle pari opportunità per uomini e donne, la procedura legislativa ordinaria adottata non richiede l'unanimità, ma unicamente la maggioranza qualificata.

La discriminazione sulla base della nazionalità è sempre stata proibita dai trattati dell'Unione, oltre che la discriminazione sulla base del sesso nel contesto dell'occupazione. Gli altri motivi di discriminazione sono stati menzionati per la prima volta nel 1997, con la sottoscrizione del trattato di Amsterdam.

Nel 2000 sono state adottate due direttive:

- la direttiva sulla parità di trattamento in materia di occupazione, che vieta la discriminazione basata su orientamento sessuale, fede religiosa, età e disabilità nel settore dell'occupazione;
- la direttiva sull'uguaglianza razziale che vieta la discriminazione basata sulla razza o sull'etnicità, sempre nel settore dell'occupazione, ma anche in settori quali istruzione, previdenza sociale, compresi sicurezza sociale e assistenza sanitaria, prestazioni sociali, accesso e fornitura di beni e servizi. La legislazione dell'Unione tutela inoltre le persone

contro la discriminazione basata sul sesso nei settori di cui sopra, a eccezione del settore dell'istruzione.

Nel 2009, il trattato di Lisbona ha introdotto una clausola cosiddetta orizzontale volta a integrare la lotta contro le discriminazioni in tutte le politiche e le azioni dell'Unione (articolo 10 del TFUE). Secondo la clausola, tutte le persone possono esercitare il proprio diritto di ricorso in caso di discriminazione diretta o indiretta, ovvero in casi di trattamento differente in un contesto equiparabile senza una giustificazione oggettiva e legittima. Le vittime di discriminazione possono inoltre richiedere assistenza agli organismi nazionali per la parità, ossia enti pubblici presenti sul territorio dell'Unione europea che si adoperano per la promozione della parità e della lotta contro la discriminazione.

Il tema della equità negli ordinamenti giuridici e vastissimo e chiaramente riguarda anche lo Stato italiano. Il docente interessato a utilizzare questo testo a fini formativi può assegnare agli studenti lo sviluppo di tesine per specifici approfondimenti, ad esempio per indagare gli ambiti della discriminazione coperti e non coperti da leggi nei diversi paesi.

Capitolo 6 – Quali principi etici chiediamo di rispettate a un modello basato sul Machine learning?

Con lo scopo di fondare eticamente i modelli introdotti in precedenza, e le tecniche di Machine learning che ne permettono la costruzione e utilizzo, in questo capitolo facciamo riferimento ad alcuni principi emessi negli ultimi anni da comunità scientifiche, organizzazioni internazionali, confessioni religiose. Parleremo sia dei principi etici validi in generale per la Intelligenza Artificiale che dei principi per i modelli basati su tecniche di Machine learning.

1. Principi della Conferenza di Asilomar

Vediamo in Figura 6.1 i principi definiti nella Conferenza di Asilomar del 2017. La sigla IA significa Intelligenza Artificiale.

Sicurezza – I sistemi di Intelligenza Artificiale devono essere sicuri in tutto il loro ciclo di utilizzazione
Trasparenza negli errori – Se un Sistema di IA causa un danno, deve essere sempre possibile accertare le cause.
Trasparenza gudiziaria - Ogni coinvolgimento di un sistema di IA in una decisione di natura giudiziaria deve fornire una spiegazione soddisfacente e verificabile dalla autorità umana competente in materia
Responsabilità – I progettisti e realizzatori di sistemi di IA avanzati sono figure primarie (stakeholders) nelle implicazioni morali del loro uso proprio e improprio, e nelle azioni svolte, con la responsabilità di individuare le loro implicazioni.
Allineamento nei valori etici – I sistemi di IA caratterizzati da elevata autonomia dovrebbero essere progettati in modo da assicurare il loro allineamento con i valori della società in cui operano.
Allineamento con i valori della società - I sistemi di IA dovrebbero essere progettati così da essere compatibili con gli ideali umani di dignità, diritti, libertà e diversità culturale.
Privacy personale - Le persone devono avere il diritto di accedere, gestire e controllare l'uso dei dati che esse generano.
Libertà e Privacy – L'applicazione della IA a dati personali non deve mai ridurre o limitare la libertà reale o percepita degli esseri umani.
Benefici condivisi – Le tecnologie della AI dovrebbero fornire benefici e potenziare la qualità della vita di quante più persone possibile.
Prosperità economica – La prosperità economica creata dalla IA dovrebbe essere condivisa così da essere di beneficio della intera umanità.
Controllo sociale - Le società dovrebbero poter decidere come e se delegare decisioni ai sistemi di AI, orientandole a raggiungere obiettivi da esse stesse definiti.
Non sovvertimento – Il potere conferito ai sistemi di IA dovrebbe rispettare e migliorare, piuttosto che sovvertire, i processi sociali e civili su cui è basato il benessere e la qualità della vita della società.
Rilevanza – I sistemi di IA dovrebbero rappresentare un profondo cambiamento nella vita sulla Terra, e dovrebbero essere pianificati strategicamente e governati con cura e risorse proporzionate.
Rischi - I rischi posti dai sistemi di IA, specialmente i rischi con effetto catastrofico e con impatto sulla esistenza del genere umano, devono essere soggetti a misure di pianificazione e mitigazione commisurati all'impatto atteso.

Figura 6.1 – Principi della Conferenza di Asilomar, 2017

Come si vede, i principi fanno riferimento a tutte le aree della Intelligenza Artificiale, e lo fanno in termini molto generali, senza entrare nel merito di specifici ambiti (per esempio il

Machine learning) o definire norme e procedure per il raggiungimento degli obiettivi definiti nei principi.

6.2 Risoluzioni e documenti della Unione Europea per una Intelligenza Artificiale degna di fiducia

L'Unione Europea ha emesso diverse risoluzioni e dichiarazioni in tema di uso di quantità massive di dati (i cosiddetti Big data), Intelligenza Aritificiale, Robotica e Sistemi Autonomi.

Il Parlamento Europeo nel 2017 ha emesso una risoluzione in tema di uso di grandi quantità di dati (i cosiddetti big data) e di Intelligenza artificiale, i cui principi guida sono riportati in Figura 6.2. Emerge, con riferimento alla attività delle forze dell'ordine, la questione etica della equità e non discriminazione.

Per utilizzare i big data per scopi commerciali e nel settore pubblico, il Parlamento europeo invita la Commissione Europea, gli Stati membri e le autorità di protezione dei dati a identificare e adottare tutte le misure possibili per ridurre al minimo la discriminazione algoritmica e sviluppare un forte e quadro etico comune per il trasparente trattamento dei dati personali e i processi decisionali così da regolare l'utilizzo dei dati e la continua applicazione del diritto dell'Unione.

Quando si tratta di utilizzare i big data per le forze dell'ordine il Parlamento "avverte che […] è necessaria la massima cautela al fine di prevenire illecite discriminazioni verso specifiche persone o gruppi di persone in riferimento a etnia, colore, origine etnica o sociale, caratteristiche genetiche, lingua, espressione o identità di genere, orientamento sessuale, stato di residenza, salute o appartenenza a minoranze."

Inoltre, il Parlamento europeo evidenzia la necessità di principi etici in materia di sviluppo della robotica e dell'intelligenza artificiale per uso civile. Sottolinea che un quadro etico dovrebbe essere "basato su […] il principi e valori sanciti dall'articolo 2 del nel Trattato sull'Unione Europea e nella Carta dei Diritti Fondamentali, quali la dignità umana, uguaglianza, giustizia ed equità, non discriminazione, consenso informato, vita privata e familiare e dati protezione".

Figura 6.2 - Dalla risoluzione del Parlamento Europeo del 2017
in tema di big data e Intelligenza artificiale

Il Gruppo Europeo sull'Etica nella ricerca nella Scienza e nuove tecnologie ha emanato nel 2018 linee guida, riportate in sintesi in Figura 6.3, che ricordano da vicino quanto espresso nella Conferenza di Asilomar.

Dignità umana – Il principio della dignità umana, consistente nel riconoscimento della condizione inerente gli esseri umani di essere degni di rispetto, non deve essere violato dalle tecnologie cosiddette *autonome*.
Autonomia – Il principio della autonomia si traduce nel riconoscimento della responsabilità umana esercitata dai sistemi autonomi, che non devono mai compromettere la libertà degli esseri umani di stabilire le proprie norme e di vivere nel loro rispetto.
Responsabilità – Il principio della responsabilità deve essere fondamentale nella ricerca sulla IA e nelle sue applicazioni. I sistemi autonomi dovrebbero essere sviluppati ed utilizzati solo in modo che servano il bene sociale e ambientale globale, come determinato da esiti di processi democratici.
Giustizia, equità e solidarietà - L'IA dovrebbe contribuire alla giustizia globale e alla parità di accesso ai benefici e vantaggi che l'IA, la robotica e i sistemi autonomi possono apportare. Distorsioni discriminatorie nei dati utilizzati per addestrare e gestire i sistemi di IA dovrebbero essere prevenute, individuate, segnalate, neutralizzate nella fase iniziale dello sviluppo e utilizzo.
Democrazia – Le decisioni sulla regolazione dello sviluppo e le applicazioni della IA dovrebbero essere il risultato di dibattito democratico e coinvolgimento sociale.
Stato di diritto e responsabilità - Lo stato di diritto, l'accesso alla giustizia e il diritto al risarcimento e a un processo equo costituiscono il quadro giuridico necessario per garantire il rispetto dei diritti umani.
Sicurezza, incolumità, integrità fisica e mentale - La sicurezza e la protezione nell'ambito dei sistemi autonomi si concretizzano in tre forme: (1) sicurezza esterna per l'ambiente e per gli utenti, (2) affidabilità e robustezza interna, ad es. contro l'hacking e (3) sicurezza emotiva e mentale nell'interazione tra l'essere umano e la macchina.
Protezione dei dati e privacy - In un'era di raccolta onnipresente e massiccia di dati per mezzo delle tecnologie digitali della comunicazione, il diritto alla protezione dei dati personali e il diritto al rispetto della privacy deve essere garantito come fondamentale.
Sostenibilità - La tecnologia AI deve essere in linea con la responsabilità umana di garantire le precondizioni fondamentali per la vita sul nostro pianeta, la continua prosperità dell'umanità e la conservazione dell'ambiente per le generazioni future.

Figura 6.3 - Principi definiti nella Dichiarazione su Intelligenza Aritificiale, Robotica e Sistemi Autonomi del 2018

6.3 – Il Gruppo G20

Il Gruppo 20 (o G20) è un forum dei leader, dei ministri delle Finanze e dei governatori delle banche centrali, creato nel 1999 dopo una successione di crisi finanziarie per favorire la dimensione internazionale della economia e la concertazione, tenendo conto delle nuove economie in sviluppo. Di esso fanno parte l'Unione Europea e 19 paesi tra i più industrializzati del mondo.

Il G20 emette periodicamente report di policy sui principali temi politici, economici, sociali ed etici comuni ai paesi aderenti al forum.

In Figura 6.4 mostro i principi del G20 sulla Intelligenza Artificiale human centric, ripresi dal documento in nota[8].

1. Fondare l'IA centrata sulla capacità degli esseri umani di governarla, sulla trasparenza, spiegabilità, equità, giustizia, inclusività, sostenibilità e formazione, combinando aspetti tecnologici e considerazioni filosofiche. Adottare un quadro giuridico sui diritti umani fondamentali.
2. Interpretare i sistemi di IA come supporto al processo decisionale umano, non come sostituto.
3. Non riconoscere le macchine come agenti morali e non dare loro una personalità o identità digitale.
4. Richiedere spiegabilità e trasparenza nei sistemi di IA.
4. Definire metriche per valutare l'impatto dell'IA su equità e giustizia sociale e prevedere di migliorare tali metriche.
6. Applicare a tutte le decisioni riguardanti l'IA un approccio di consultazione di tutti i portatori di interessi.
7. Misurare l'impatto dell'IA sull'ambiente. Considerare il benessere delle generazioni future al momento di decidere iniziative, incentivi, finanziamenti, e politiche che riguardino tecnologie di IA.
8. Includere l'etica dei dati e della tecnologia nei curricula scientifici. Espandere le iniziative di apprendimento permanente.
9. Creare attività di alfabetizzazione AI per i cittadini.
10. Creare comitati etici indipendenti e multi-disciplinari in ogni Paese.
11. Nella regolazione della IA, imporre condizioni sull'uso della IA (e non sulla IA in astratto) e adottare un approccio per cui le regole di uno specifico paese si applicano a chiunque opera in quel paese.

Figura 6.4 – Principi del G20 sulla Intelligenza Artificiale human centric

6.4. La Call for AI Ethics

Al fine di sostenere un approccio etico all'Intelligenza Artificiale e promuovere tra organizzazioni, governi e istituzioni un senso di responsabilità condivisa, la Pontificia Accademia per la Vita, Microsoft, IBM, la FAO, il Governo italiano, firmarono il 20 febbraio 2021 la Call for AI Ethics con l'obiettivo di garantire un futuro in cui l'innovazione digitale e il progresso tecnologico siano al servizio del genio e della creatività umana e non la loro graduale sostituzione, vedi in Figura 6.5 il momento della firma del documento finale.

Figura 6.5 – La firma della dichiarazione finale dell'incontro Rome Call by AI Ethics

[8] Carlo Casalone, Luciano Floridi, Laura Palazzani, Renzo Pegoraro, Francesca Rossi, Roberto Villa, Human Centric AI: From Principles to Actionable and Shared Policies, G20 Insights, September 2021.

I firmatari della Call for AI Ethics espressero il desiderio di lavorare insieme, a livello nazionale e internazionale, per promuovere una "algor-etica", ovvero lo sviluppo e l'utilizzo dell'Intelligenza Artificiale secondo i principi riportati in Figura 6.6.

1. Trasparenza: gli algoritmi dell'Intelligenza Artificiale dovrebbero essere spiegabili, comprensibili;
2. Inclusione: le necessità degli esseri umani devono essere prese in considerazione così che chiunque ne possa trarre beneficio e alle persone possano essere offerte le migliori condizioni possibili per esprimersi e svilupparsi;
3. Responsabilità: chi progetta e sviluppa I sistemi di intelligenza artificiale deve farlo con responsabilità e rendendo trasparenti le scelte.
4. Imparzialità o Equità: gli algoritmi della Intelligenza Artificiale non devono agire in accordo a discriminazioni, salvaguardando in questo modo l'equità e la dignità umana.
5. Affidabilità: I sistemi di Intelligenza Artificiale devono operare senza errori o guasti.
6. Sicurezza e privacy: I sistemi di Intelligenza Artificiale devono operare in modo sicuro e rispettando la privacy degli utenti.

Figura 6.6 – I principi etici per la Intelligenza Artificiale stabiliti dalla Call for AI Ethics

Come si vede, il documento esprime uno sforzo per focalizzare i principi etici su aspetti concreti dell'etica, ed in particolare su tre principi che sono investigati nel seguito, la *Equità*, la Affidabilità, chiamata in questo libro *Accuratezza*, e la Trasparenza, chiamata in questo libro *Spiegabilità*; queste tre qualità sono investigate con riferimento alle tecniche di Machine learning e ai modelli classificatori, predittivi e decisionali che il Machine learning realizza.

Cominciamo a intuire, anche se avremo bisogno di arrivare al Capitolo 8 per una completa comprensione, che i modelli classificatori, predittivi, decisionali possono produrre discriminazioni su esseri umani o gruppi sociali (Principio 4, *Equità*), e producono soluzioni affette da errori (Principio 5. *Accuratezza*). Vedremo poi che è una esigenza rilevante comprendere quale processo abbia seguito il modello per formulare la classificazione, la previsione, la decisione (Principio 1. *Spiegabilità*), e dedicheremo molte delle considerazioni del Capitolo 10 sui metodi di mitigazione disponibili per i progettisti dei modelli e l'intera Società nel creare antidoti ad un uso dei modelli che violi i principi di Accuratezza e Equità.

Bene, ora sappiamo quale è la nostra meta, procediamo!

Sono proprio curioso, è come se finora fossimo ancora all'antipasto!

Capitolo 7 - Cosa chiediamo ai modelli basati su tecniche di Machine learning? L'Accuratezza

In questo capitolo affrontiamo la questione della accuratezza nei modelli basati su tecniche di Machine learning guidate da esempi. Anche qui partiamo da un esempio, che riprende l'esempio finale del Capitolo 4.

Consideriamo un albero di decisione molto semplice, in cui il livello di rischio e la decisione se concedere la libertà provvisoria è valutato partendo da una sola caratteristica dei detenuti, il loro *genere*. Supponiamo che nel passato donne e uomini abbiano commesso o non commesso recidiva e quindi siano a rischio alto o basso nelle proporzioni descritte in Figura 7.1. Riportiamo anche la decisione finale sulla attribuzione tra rischio basso e alto nell'albero di decisione.

Genere/Rischio →	Rischio basso	Rischio alto	Livello di rischio
Donne	1.200	800	Basso
Uomini	500	1.500	Alto

Figura 7.1 - L'unica cosa che conosciamo è il genere…..

A partire dai numeri di Figura 7.1 possiamo costruire l'albero di decisione di Figura 7.2; l'albero ci dice che se il detenuto è una donna viene concessa la libertà provvisoria, se è un uomo no, e questo perché tra le donne è superiore la percentuale di coloro che non hanno commesso recidiva, mentre per gli uomini accade il contrario.

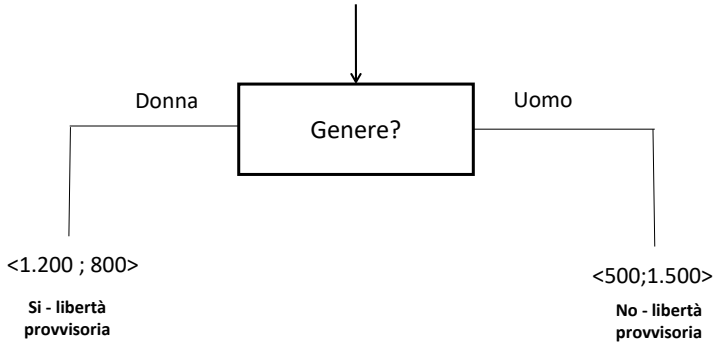


Figura 7.2 - Albero di decisione nel caso di Figura 7.1

Applichiamo ora il modello predittivo *alle stesse persone* considerate per costruire l'albero di decisione.

Ma come puoi applicarlo alle stesse persone che ti hanno permesso di costruirlo?

Lo hai fatto anche tu nella prima delle tue domande del Capitolo 4! Comunque: quelle persone mi hanno permesso di costruire l'albero iniziale, ma poi, nell'albero di decisione finale, il rischio *è stato calcolato a maggioranza*…

Sì, non ci avevo pensato: ci sono due alberi di decisione, quello che corrisponde alla parte azzurra di Figura 7.1 e quello che corrisponde alla parte gialla, questo è l'albero di decisione finale!

Ebbene, confrontando le due parti di Figura 7.1, arriviamo a conclusioni piuttosto inquietanti: ci sono infatti 800 donne che sono ad alto rischio, ma a cui viene attribuito rischio basso e a cui, di conseguenza viene concessa la libertà provvisoria; queste donne cotituiscono una minoranza all'interno della comunità delle donne, in cui la maggioranza si è comportata correttamente e non ha commesso nuovi reati.

Parallelamente, ci sono 500 uomini che pur non avendo commesso recidiva, hanno associato rischio alto e non vengono liberati, perché appartengono a una comunità, quella degli uomini, che in grande maggioranza ha commesso recidiva ed è dunque a rischio alto.

Insomma, nel caso delle 800 donne e dei 500 uomini percepiamo una situazione di *iniquità*; magari, a seconda delle nostre sensibilità, percepiamo forse maggiore iniquità nel caso degli uomini piuttosto che nel caso delle donne, perché in fondo le donne *vengono liberate*, anche se erano state considerate ad alto rischio, mentre gli uomini *non vengono liberati*, anche se erano stati considerati a basso rischio.

7.1 Gli errori di classificazione: i veri e falsi positivi, i veri e falsi negativi

E' possibile essere più precisi nella definizione dei precedenti aspetti guardando la Figura 7.3. Consideriamo un modello predittivo *applicato alle stesse persone con cui è stato costruito l'albero di decisione.*

Concentriamoci prima sulla figura a sinistra, che fa riferimento alle persone in generale, senza distinguerle tra donne e uomini.

Vengono definite *vere positive* le persone cui corrisponde alto rischio di recidiva e a cui è assegnato dall'albero di decisione alto rischio, corrispondente perciò alla decisione di non liberarle; le vere positive sono, insomma, le persone che, secondo il criterio di decisione adottato, *è giusto non liberare*.

Vengono definite *vere negative* le persone cui è assegnato basso rischio di e a cui è assegnato dall'albero di decisione basso rischio, corrispondente perciò alla decisione di liberarle; le vere negative sono, insomma, le persone che, secondo il criterio di decisione adottato, *è giusto liberare*.

Vengono definite *false positive* le persone cui è assegnato basso rischio di recidiva e a cui è assegnato dall'albero di decisione alto rischio, e che non vengono liberate; le false positive sono, insomma, le persone che, secondo il criterio di decisione adottato, *non è giusto non liberare.*

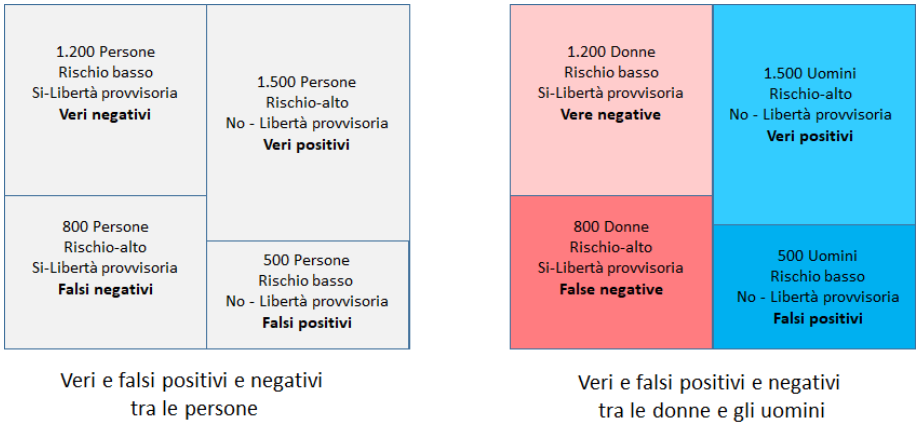


Figura 7.3 – Veri e falsi positivi, veri e falsi negativi tra le persone, le donne, gli uomini

Vengono definite infine *false negative* le persone con alto rischio di recidiva e a cui è assegnato dall'albero di decisione basso rischio, e che vengono perciò liberate; le vere negative sono, insomma, le persone che, secondo il criterio di decisione adottato, *non è giusto liberare*.

Vediamo in Figura 7.4 tre ambiti in cui sono applicati modelli classificatori (Test Covid) e predittivi (libertà provvisoria, prestiti bancari), con esemplificazione dei veri e falsi positivi e negativi.

	Veri positivi	**Falsi positivi**	**Veri negativi**	**Falsi negativi**
Test Covid	Le persone che risultano positive e che hanno il Covid	Le persone che risultano positive e che non hanno il Covid	Le persone che risultano negative e che non hanno il Covid	Le persone che risultano negative e che hanno il Covid
Prestito bancario	Le persone classificate ad alto rischio di non restituire il prestito che sono effettivamente tali	Le persone classificate ad alto rischio di non restituire il prestito che non lo sono	Le persone classificate a basso rischio di non restituire il prestito che sono effettivamente tali	Le persone classificate a basso rischio di non restituire il prestito che non lo sono
Concessione della libertà provvisoria	Le persone classificate ad alto rischio di recidiva che sono effettivamente tali	Le persone classificate ad alto rischio di recidiva che non lo sono	Le persone classificate a basso rischio di recidiva che sono effettivamente tali	Le persone classificate a basso rischio di recidiva che non lo sono

Figura 7.4 - Definizioni di riferimento ed esempi per veri e falsi positivi e negativi

7.3 Le misure di qualità: Precisione e Recall

E ora definiamo le tre misure più utilizzate per esprimere la accuratezza di un modello classificatorio o predittivo basato su tecniche di Machine learning da esempi.

L'accuratezza è definita come:

Accuratezza = **((numero di,** sarà omesso nel seguito**…) veri positivi + veri negativi)/ (veri positivi + falsi positivi + veri negativi + falsi negativi)**[9]

Quindi la accuratezza esprime il numero di elementi cui è attribuita una classificazione (es. ha o non ha il Covid) o previsione (è ad alto o basso rischio di recidiva) sbagliata, sul totale degli elementi.
La precisione è definita come

Precisione = veri positivi / (veri positivi + falsi positivi)

La precisione misura, nel nostro esempio, la percentuale delle persone che è giusto non liberare *tra tutte le persone che non vengono liberate*, inclusi, quindi, i falsi positivi. Al contrario della accuratezza, si concentra perciò sui positivi veri, misurandone la percentuale rispetto a tutti i positivi, veri e falsi.
La recall (tradotta spesso con recupero) è

Recall = veri positivi / (veri positivi + falsi negativi)

La recall misura nel nostro esempio la percentuale delle persone che è giusto non liberare tra tutte le persone che non dovrebbero essere liberate, inclusi, quindi i falsi negativi che vengono liberati pur essendo ad alto rischio, e che quindi non dovrebbero essere liberati.

Per comprendere meglio i concetti di precisione e recall, ci possiamo aiutare con il diagramma di Figura 7.5. Qui e nel seguito i pallini grigi rappresentano, nel nostro esempio, le persone con alto rischio di recidiva e perciò *è giusto non liberare,* mentre i pallini Bianchi rappresentano le persone con basso rischio di recidiva e perciò *è giusto liberare. Osservando la parte sinistra della figura, i pallini nel rettangolo marcato con cornice blu, cioè le persone che effettivamente non vengono liberate,* si dividono tra quelli relativi alle persone che è giusto non liberare (rettangolo rosa chiaro) e quelli che è ingiusto non liberare (rettangolo rosa scuro). La proporzione tra i pallini grigi nel rettangolo rosa chiaro e tutti i pallini nei due rettangoli rosa è la Precisione.

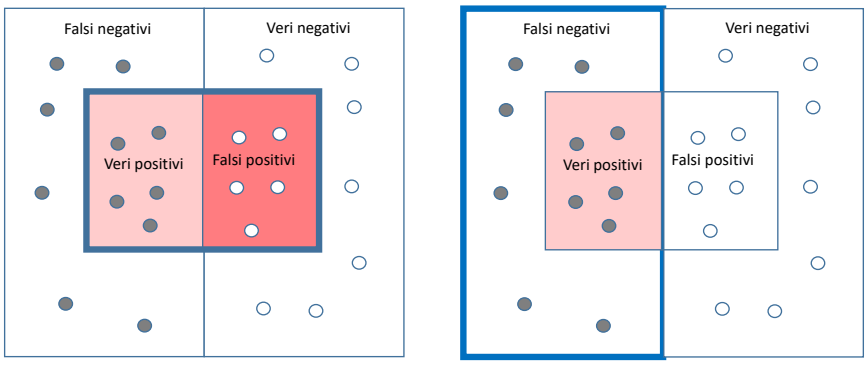


Figura 7.5 – Rappresentazione grafica della Precisione e della Recall

[9] Ricordo ancora che in queste formule abbiamo assunto che tutti gli elementi nei veri e falsi positivi e negativi abbiano lo stesso peso. Si può anche assegnare pesi superiori, ad esempio, ai falsi negativi, perché hanno una importanza maggiore nei risultati del modello, come accade nei test di efficacia di un vaccino.

Nel rettangolo a destra marcato in blu i pallini grigi nel rettangolo rosa sono le persone che è giusto non liberare, mentre le persone nella parte bianca sono le persone che è ingiusto liberare, e quindi la proporzione tra i primi pallini e tutti i pallini corrisponde alla Recall.

Per dirla in termini più immediati, la precisione misura *quanto indoviniamo* nel *non liberare* le persone, mentre la recall misura *quanto indoviniamo* nel *non liberare quelli che non dovrebbero essere liberati*.

7.4 Specificità e sensibilità

Sì, dopo essermi riletto con attenzione due volte le precedenti definizioni e dopo aver guardato le figure, credo di aver capito. Però volevo chiedere una cosa: durante tutto il periodo della fase acuta della epidemia Covid, quando si è parlato di test che permettevano di diagnosticare i positivi al virus, si è fatto un gran parlare di Specificità e Sensibilità, che relazione c'è tra queste due caratteristiche e la Precisione e il Recall?

Allora, specificità e sensitività sono definite nel modo seguente

Specificità = (numero di) veri negativi / (falsi positivi + veri negativi)

Sensitività = veri positivi / (veri positivi + falsi negativi)

in cui qui con *negativi* si intende *negativi al test*. La specificità è dunque il rapporto tra le persone negative al test che non sono malate di Covid e il totale delle persone *che non sono malate di Covid*, in cui sono incluse le persone che non sono malate di Covid ma sono risultate positive al test.

Insomma la specificità misura quanto il test indovina i veri negativi tra tutti i negativi. La specificità è massima quando i diagnosticati negativi sono tutti i negativi. Se un test ha un'alta specificità, allora è basso il rischio di falsi positivi, cioè di soggetti che pur presentando valori anomali non sono affetti dalla patologia che si sta ricercando.

La sensibilità di un esame diagnostico coincide con la Recall, è dunque, in un contesto di diagnosi di malattia, la capacità di identificare correttamente i soggetti affetti dalla malattia o dalla condizione che ci si propone di individuare. Se un test ha un'ottima sensibilità, allora è basso il rischio di falsi negativi, cioè di soggetti che pur presentando valori normali sono comunque affetti dalla patologia o dalla condizione che si sta ricercando.

Vedi in Figura 7.6 la rappresentazione grafica della Specificità e della Sensibilità.

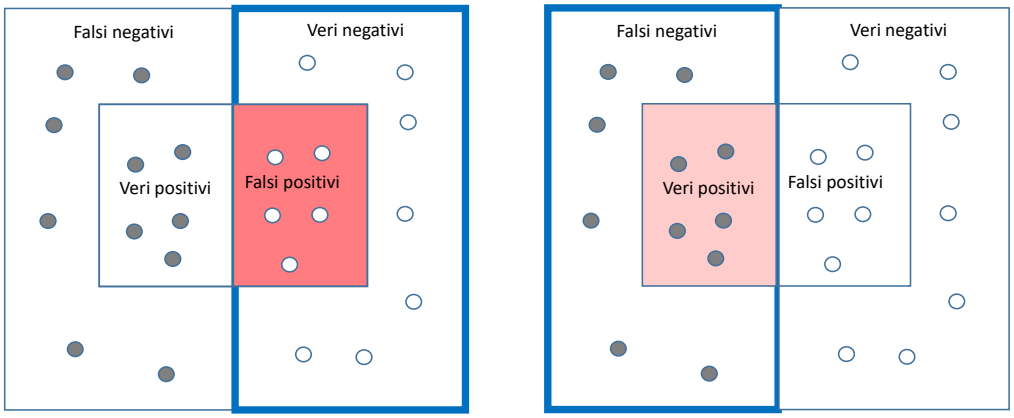


Figura 7.6 – Rappresentazione grafica della Specificità e Sensibilità

7.5 Come misurare le dimensioni di Accuratezza, Precisione e Recall

Facciamo il punto. Tornando agli esempi sulla concessione della libertà provvisoria, ho avuto una sensazione strana, come dire, che "te la suoni e te la canti". Mi spiego meglio: tu applichi il modello predittivo del rischio di recidiva agli stessi soggetti che ti sono serviti per generare l'albero di decisione e il modello predittivo…Ma che senso ha applicare al passato un modello predittivo sui casi che l'hanno generato?

Capisco la stranezza, ma un senso ce l'ha! E' vero che ho creato il modello predittivo su casi del passato, ma è anche vero che nell'esempio di Figura 7.1, delle 2.000 donne del passato solo 1.200 non avevano commesso recidiva, mentre le altre 800 *pur avendola commessa, nella applicazione retroattiva del metodo, sarebbero state liberate*! E una dinamica simile avviene per gli uomini: 500 di loro non hanno commesso recidiva, ma *nella applicazione retroattiva del metodo*, non avrebbero avuto concessa, ingiustamente, la libertà!

Insomma, i casi del passato ci sono utili per costruire il modello predittivo, ma noi possiamo applicare il modello predittivo anche ai casi del passato per capire quali sarebbero stati i veri e falsi positivi e negativi, ciò è perfettamente lecito!

In realtà, l'esempio che abbiamo visto è molto utile per farci capire una questione della massima importanza: una volta costruito il modello predittivo a partire dai dati del passato, *dobbiamo sempre verificare quanto il modello predittivo, usato per decidere sulla libertà provvisoria o per diagnosticare una malattia*, o per qualunque altro uso sia caratterizzato, a seconda dei casi, da elevata precisione e recall, oppure da sensibilità e specificità.

E questo può farsi se usiamo la seguente procedura, che mostriamo per mezzo della Figura 7.7.

Quando noi utilizziamo un insieme di dati per generare un modello predittivo, abbiamo la esigenza di verificare il modello sulle misure di qualità che abbiamo definito poco fa. Per fare ciò, possiamo decidere di *separare l'insieme dei dati in due parti*, di cui la prima viene usata

per l'addestramento, mentre la seconda è tenuta da parte per verificare quanto il modello rispetti le misure di qualità che ci paiono più utili tra accuratezza, precisione, sensibilità (o recall) e specificità. Chiameremo il primo insieme di dati *dati di addestramento*, e il secondo *dati di test*.

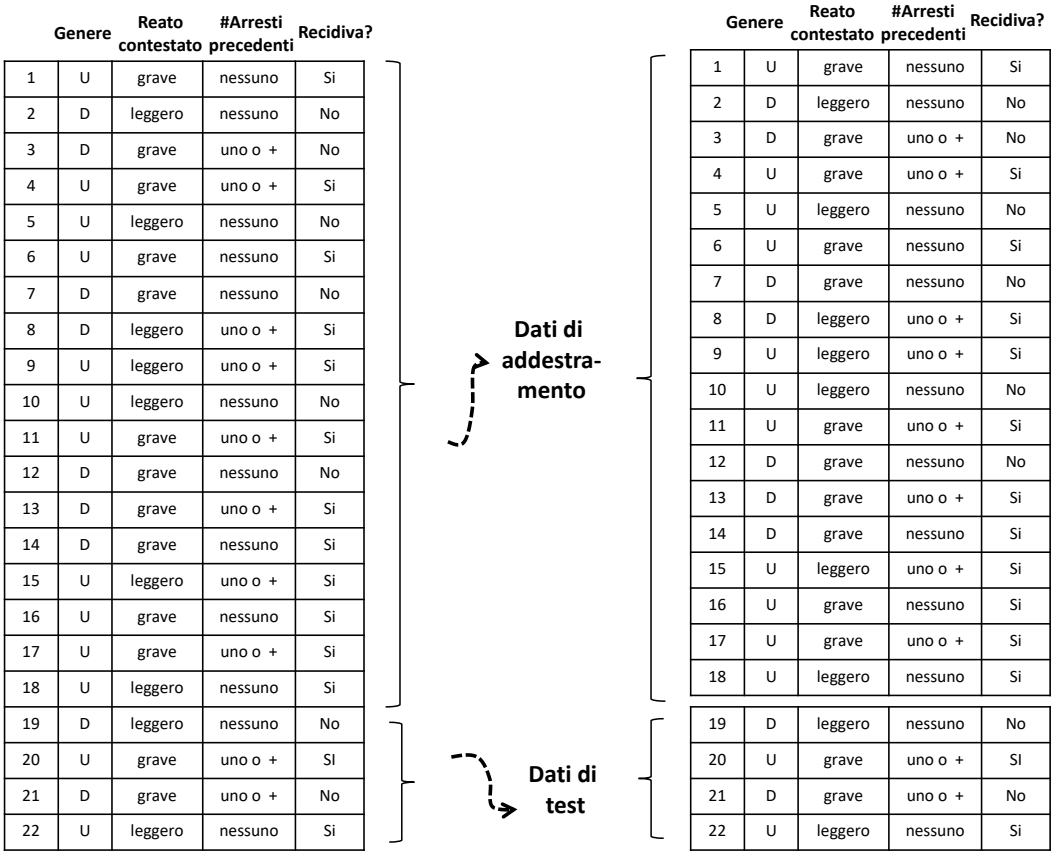


Figura 7.7 – Separazione tra dati di addestramento e dati di test

Bisogna fare molta attenzione nello scegliere quali dati sono di addestramento e quali di test; ad esempio, non dobbiamo scegliere i dati di test in maniera *maliziosa*, cioè in modo tale che il test dia risultati con numero ridotto o al limite nessun falso positivo e falso negativo.

In fondo non sarebbe complicato: nel caso della libertà provvisoria, in cui abbiamo a disposizione la sola caratteristica costituita dal genere, potremmo selezionare solo donne che non hanno commesso recidiva e uomini che l'hanno commessa, generando così solo veri positivi e veri negativi, e ottenendo il valore massimo pari a 1 per le misure di accuratezza[10]. Questa sarebbe però una finta accuratezza, perché nel mondo reale non esistono solo donne che non commettono recidiva e solo uomini che la commettono!

Applichiamo dunque il precedente metodo ai dati che compaiono nella tabella a sinistra di Figura 7.7. Separiamo i dati nei due insiemi di addestramento e di test. Per fare in modo che i due insieme manifestino valori simili per le dimensioni di qualità, potremmo scegliere a caso

[10] Qui e nel seguito il termine accuratezza indicherà genericamente l'insieme delle misure che abbiamo chiamato accuratezza, precisione, recall

le righe per l'addestramento e per il test. Qui noi semplifichiamo, e consideriamo per l'addestramento le prime 18 righe, e per il test le rimanenti quattro.

Esercizio 7.1 – Genera l'albero di decisione partendo dai dati di addestramento in Figura 7.7. Genera poi il modello predittivo.

In Figura 7.8 ho rappresentato l'albero di decisione con la numerosità dei soggetti a basso e alto rischio per ciascuna delle foglie dell'albero; ho rappresentato mediante pallini grigi i vari elementi associati alla relativa foglia cui corrisponde. In Figura 7.9 ho rappresentato le decisioni associate alle varie foglie, in cui ho direttamente segnato in questo caso gli esiti finali, No-libertà provvisoria o Sì-libertà provvisoria.

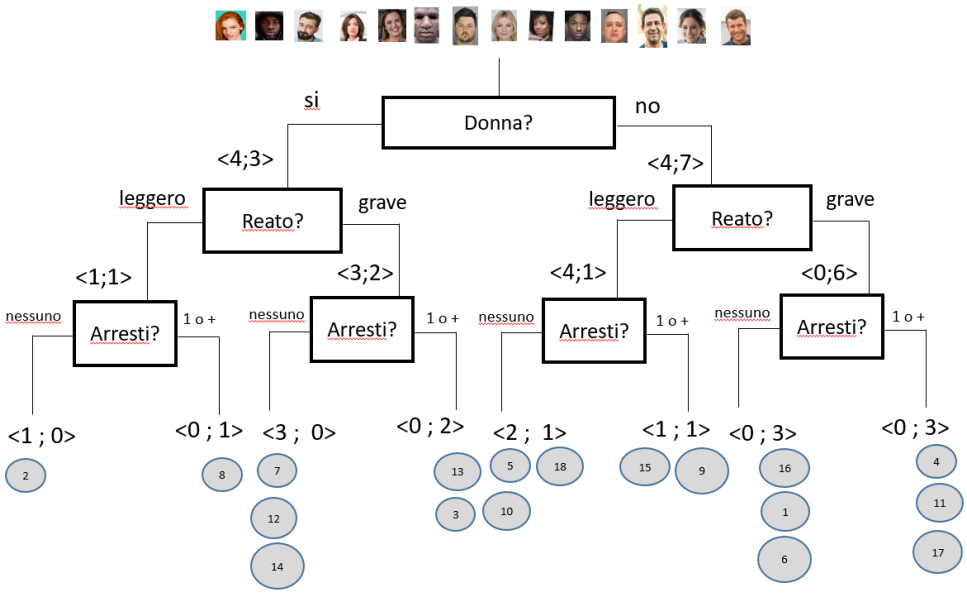


Figura 7.8 – Albero di decisione per i dati di addestramento

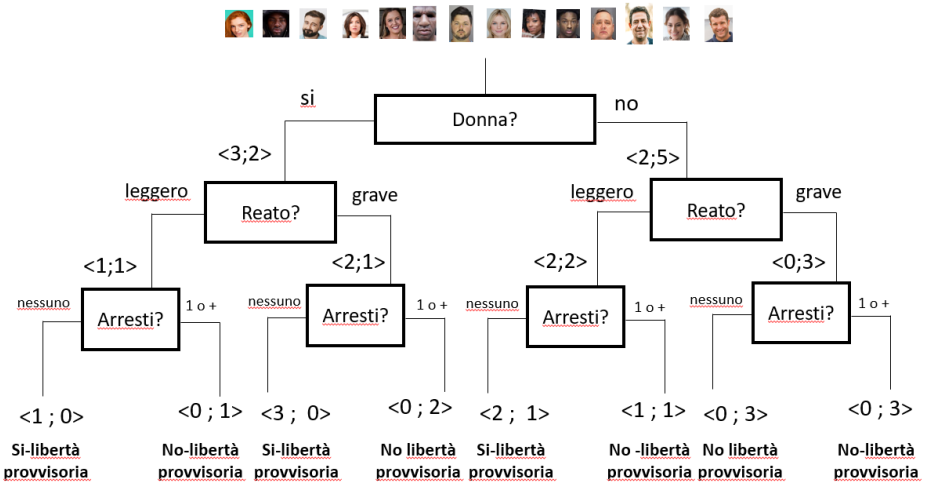


Figura 7.9 – Modello predittivo per i dati di addestramento

Esercizio 7.2 - Adesso associamo le quattro persone nei dati di test di figura 7.6 alle foglie cui corrispondono nella classificazione. Rappresentiamo anche il numero dei veri positivi, veri negativi, falsi positivi, falsi negativi. Anche in questo caso rappresentiamo l'esito finale della valutazione di rischio. La risposta qui sotto….

Ecco in Figura 7.10 la risposta all'esercizio. Ad esempio, la persona 19, le cui caratteristiche la portano ad essere a basso rischio, è una vera negativa, perché la decisione associata alla foglia è quella di concedere la libertà provvisoria. La persona 21 è a basso rischio ma non le viene concessa la libertà provvisoria, quindi è un falso positivo.

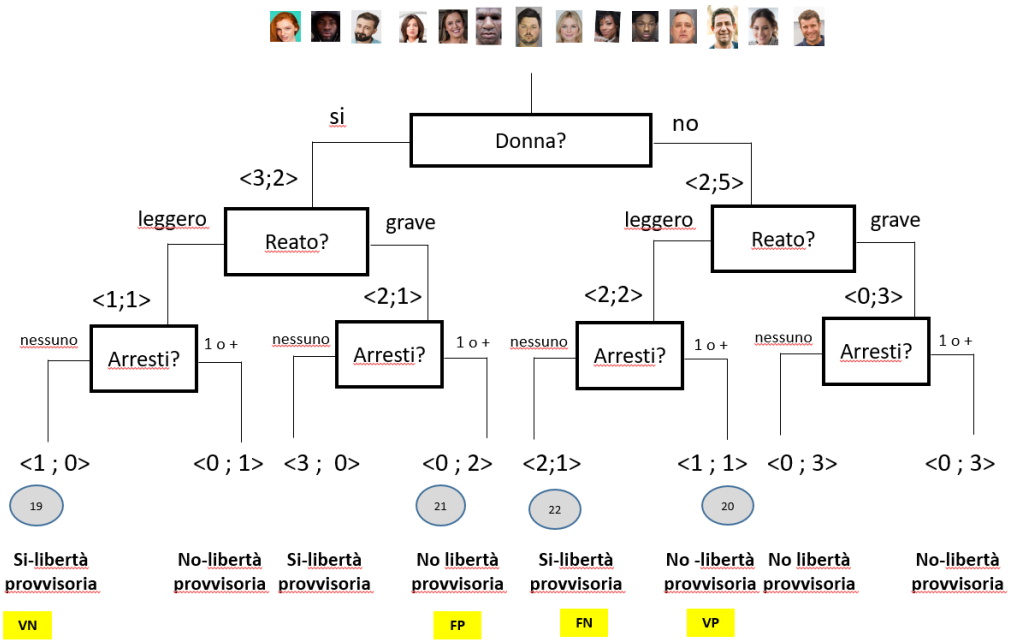


Figura 7.10 – Risposta all'esercizio

Esercizio 7.3 - Ora vi propongo di calcolare Precisione e Recall riferiti agli elementi presenti nei dati di test.

I risultati nella prossima pagina.

Per i risultati, vedi Figura 7.11.

Tutti	Positivi	Negativi
Veri	1	1
Falsi	1	1

Precisione = ½ = 50%
Recall = ½ = 50%

Figura 7.11 - Valori di Precisione e Recall per i dati di test

Per esempio, per calcolare la Precisione dobbiamo usare la formula

Precisione = (numero di) veri positivi / (veri positivi + falsi positivi) = 1 / 2 = 0,5

Ora a titolo di esercizio, potreste provare a costruire il modello predittivo del rischio e il modello predittivo per i dati della tabella della Figura 7.7, in cui puoi scegliere come dati di test, invece che gli ultimi quattro, i primi quattro. Non riporto le soluzioni.

7.6 Dagli alberi alle foreste di alberi

Uno dei principali svantaggi degli alberi di decisione è che funzionano bene sui dati di addestramento, ma non sono flessibili per fare previsioni su dati nuovi; con i dati nuovi, cioè tendono ad avere una precisione più limitata; Il vantaggio degli alberi di decisione è che sono di facile interpretazione.

L' *apprendimento ensemble* prevede l'utilizzo, invece che di un modello predittivo, di un ampio numero di modelli predittivi contemporaneamente, per migliorare le prestazioni di ciascuno di loro considerato individualmente. Questo metodo di apprendimento, in altre parole, utilizza un insieme di modelli ciascuno dei quali è poco accurato, al fine di crearne uno più preciso. Nel nostro caso, le *foreste casuali* sono un insieme di molti alberi di decisione.

Le foreste casuali combinano la semplicità degli alberi di decisione, con la flessibilità e la potenza di modelli in cui diversi alberi insieme concorrono a formare la decisione. Sebbene le foreste casuali non offrano la stessa capacità di interpretazione di un singolo albero, le loro prestazioni sono di gran lunga migliori e non dobbiamo preoccuparci così tanto della regolazione precisa dei parametri della foresta (per esempio l'ordine con cui consideriamo le caratteristiche) come facciamo con i singoli alberi.

Costruzione di una foresta casuale

La costruzione di una foresta casuale ha tre fasi principali. Vediamole.

a. Creazione di un insieme di dati di addestramento per ogni albero

Quando creiamo un unico albero predittivo, utilizziamo un unico insieme di dati di addestramento. Abbiamo detto che facendo crescere l'albero, per esempio introducendo

nuove caratteristiche caratterizzate dal alto guadagno informativo, l'albero si adatta molto bene (certe volte *troppo bene*) a questi dati di addestramento e si adatta male a nuovi dati; è come un vestito molto ben disegnato per una persona, che mal si adatta al fisico di un'altra persona. Per risolvere questo problema, possiamo *impedire all'albero di crescere molto*.

Per costruire una foresta casuale dobbiamo dunque costruire, invece che un albero, n alberi di decisione. Una prima idea per diversificare questi alberi è quella di *non addestrarli tutti con l'intero insieme di dati di addestramento*, ma con sottoinsiemi; in questi sottoinsiemi i dati di addestramento sono scelti in modo *casuale* (da qui il nome) a partire dall'insieme di dati iniziali, eventualmente duplicando alcuni elementi dell'insieme.

Nel caso della recidiva, quindi, accade in generale che la scelta causale porti *i dati di un elemento* ad *essere usati in più alberi*. Vedi in Figura 7.12 evidenziati in giallo e in verde due elementi utilizzati due volte.

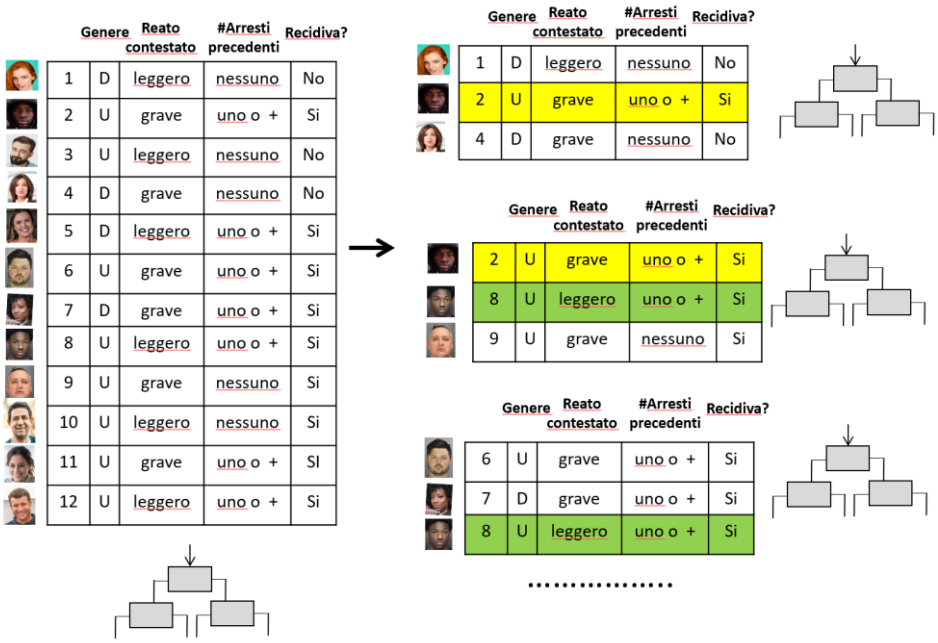


Figura 7.12 – Da un albero a tanti alberi "figli" – primo passo

Il risultato di questo primo passo è che ora addestriamo l'albero con diversi insiemi di dati, scelti casualmente. Possiamo diversificare ulteriormente l'albero iniziale, trasformandolo in una foresta, con il procedimento seguente: l'albero che opera sulle tre caratteristiche, a. genere, b. reato contestato, e c. numero di arresti viene trasformato in tre alberi, ognuno dei quali opera su due delle tre caratteristiche, vedi Figura 7.13.

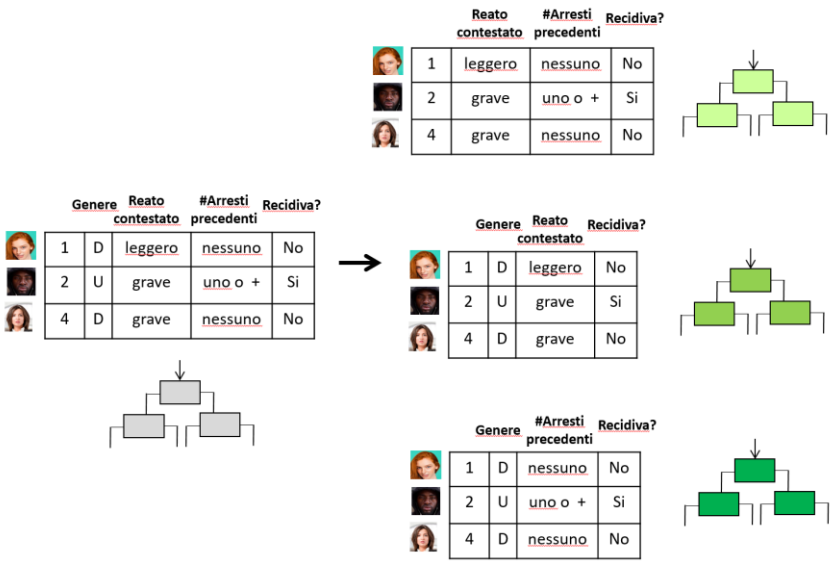


Figura 7.13 – Da un albero a tanti alberi "figli" – secondo passo

Nel caso generale, partendo da un albero con n caratteristiche, possiamo generare tanti alberi che operano su un numero k di caratteristiche scelte a caso tra le n.

Insomma, a partire da un albero che opera su un insieme di dati di addestramento e un insieme di caratteristiche, possiamo generare una foresta di alberi diversificati rispetto a quello originario, o rispetto ai dati di addestramento o rispetto alle caratteristiche su cui viene addestrato, vedi Figura 7.14 e Figura 7.15.

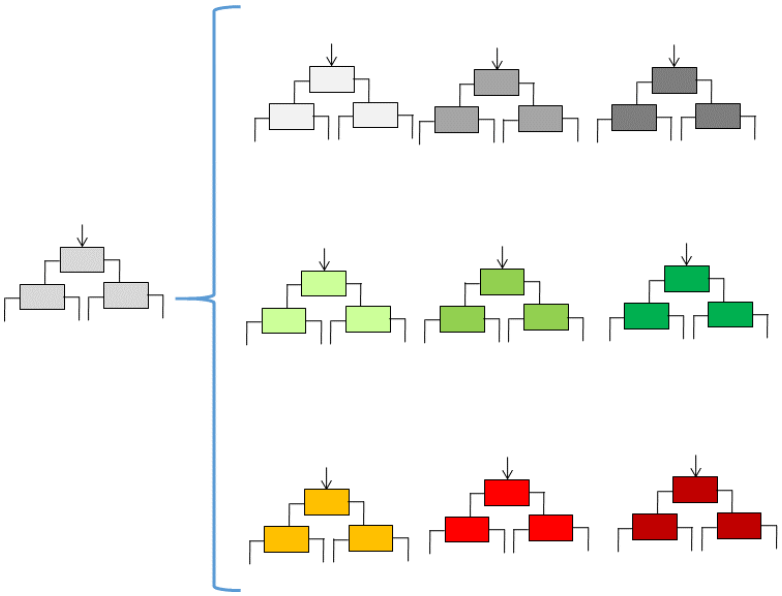


Figura 7.14 – Da un albero a tanti alberi "figli" – I due passi insieme

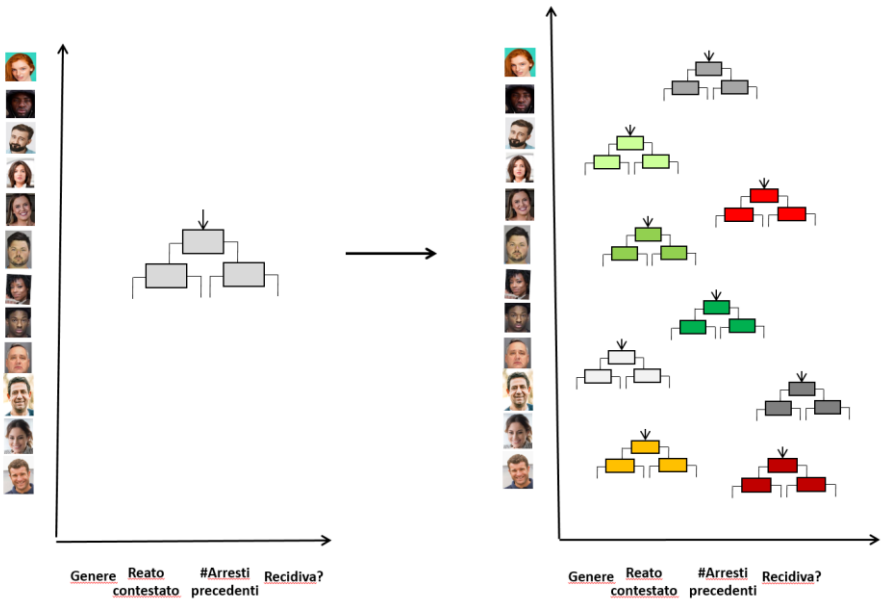


Figura 7.15 - – Da un albero a una foresta casuale di alberi

Utilizzando diversi campioni di dati per addestrare ogni singolo albero, risolviamo uno dei problemi principali che hanno gli alberi: sono molto legati ai loro dati di addestramento. Se addestriamo una foresta con molti alberi e ognuno di loro è addestrato con dati diversi, risolviamo questo problema. Se utilizziamo una porzione molto piccola dell'intero insieme di dati per addestrare ogni singolo albero, aumentiamo la casualità della foresta; maggiore è la diversità degli alberi, meglio è: riduciamo la varianza dei risultati, e otteniamo un modello con maggiore precisione.

Abbiamo imparato come costruire una foresta a partire da un primo, unico albero di decisione. Ora, possiamo ripetere il procedimento per gli N alberi creati, iterando il processo precedente. In conclusione, l'intero processo consiste nel:

1. Creare un insieme di dati di addestramento per un primo albero.
2. A partire dal primo albero, creare tanti alberi di decisione utilizzando sotto-insiemi casuali di dati di addestramento o caratteristiche.
3. Ripetere il processo di generazione di alberi del passo 2 decine e centinaia di volte per costruire una grande foresta con ampia varietà di alberi, tutti con qualche elemento di diversità rispetto agli altri. Questa *diversità* [11]è ciò che rende una foresta casuale migliore di un singolo albero di decisione.

A questo punto dobbiamo decidere come usare la foresta per prendere le decisioni. Ciò è molto semplice; dobbiamo considerare ciascuno dei nostri alberi individualmente, fare una previsione attraverso di essi, quindi ottenere una previsione complessiva e aggregata, vedi Figura 7.16.

[11] Questo è molto interessante dal punto di vista sociale: è la diversità e non la omologazione che porta ricchezza.

La generazione di un insieme di *dati di addestramento casuali* e l'utilizzo di un'aggregazione di modelli per fare una previsione sono chiamati *Bagging* e il modo in cui viene effettuata questa previsione e pesate le diverse previsioni dei singoli alberi dipende dal tipo di problema che stiamo affrontando.

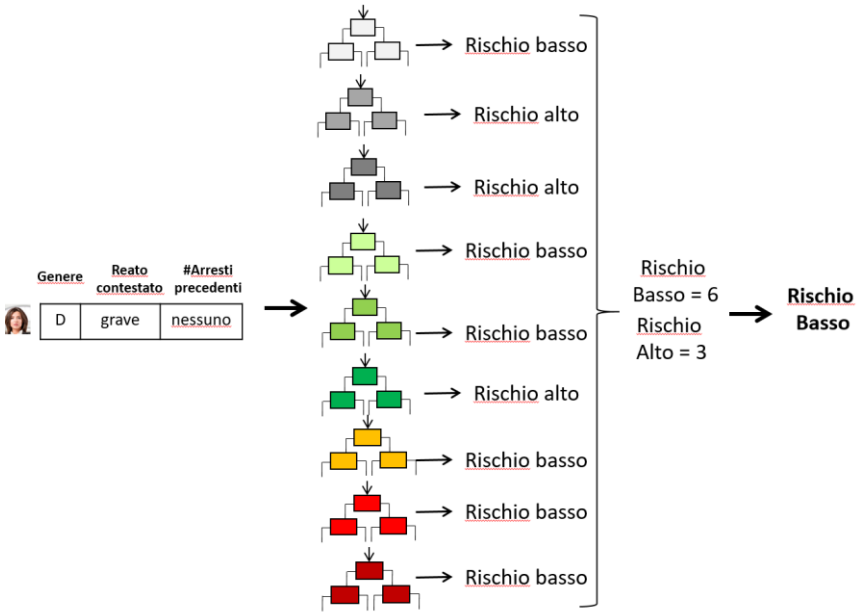


Figura 7.16 – Nuovo modello di decisione

Abbiamo dunque visto che cosa è una Foresta casuale, e come supera i problemi principali degli alberi di decisione, generando una nuova famiglia di modelli più precisi degli alberi di decisione. Allo stesso tempo, è comprensibile come a fronte di una maggiore precisione, la foresta esprima in modo meno comprensibile il processo decisionale. Torneremo su questo punto nel Capitolo 9.

7.7 Conclusioni

Quando usiamo una tecnica di Machine learning guidata da esempi per costruire un modello predittivo, siamo certi che il risultato sarà sempre affetto da errore. In questi casi, piuttosto che adottare a scatola chiusa un modello predittivo, possiamo chiedere al produttore del modello:

- quale sia il livello di accuratezza,
- su quali e quanti dati è stato prodotto il modello predittivo,
- su quali e quanti dati è stato verificato.

Insomma, meglio sapere quanto è inaccurato un modello piuttosto che non sapere niente. Pretendere di conoscere l'accuratezza e con quali dati è stata misurata, ci permette di scegliere tra i vari produttori, e di non adottare un prodotto a scatola chiusa. Inoltre, per caratterizzare l'accuratezza, abbiamo a disposizione diverse misure: l'accuratezza, la

sensibilità o Recall, la precisone, ciascuna di esse osserva uno specifico aspetto della qualità che in termini generali chiamiamo *accuratezza*.

Per esempio, se ci sottoponiamo a un test per vedere se siamo malati di una certa patologia, ad esempio il Covid, scegliere un test più costoso che ha una più alta specificità rispetto a uno meno costoso, significa minimizzare il rischio di falsi positivi, cioè di presentare valori anomali, ma non avere il Covid. Scegliere un test che ha un'alta *sensibilità (o recall)* significa minimizzare il rischio di falsi negativi, cioè di avere il Covid, ma il test non se ne accorga.

La precisone, infine osserva un diverso fenomeno, che possiamo chiamare la *capacità di previsione*, e che corrisponde a individuare i positivi (persone ad alto rischio, persone per cui viene diagnosticato il Covid, ecc.) minimizzando quelli che sono scorrettamente individuati come tali, i falsi positivi).

Capitolo 8 - Cosa chiediamo ai modelli basati su tecniche di Machine learning? La Equità, le Equità

8.1 La denuncia di ProPublica: gli Afro-Americani sono discriminati!

Nel 2016 ProPublica[12], un giornale on line che effettua inchieste di giornalismo investigativo, pubblicò una analisi sullo strumento Compas in cui affermò (vedi Figura 8.1) che lo strumento era "biased against blacks".

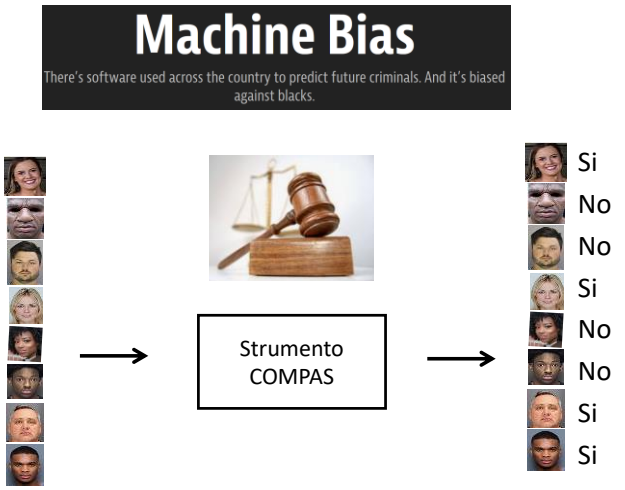


Figura 8.1 – Lo strumento Compas

ProPublica circostanziava questa sua denuncia sulla base di una analisi storica condotta su detenuti della contea di Broward, Florida, che si concentrò sui detenuti che avevano ottenuto la libertà provvisoria, e che erano stati riarrestati nei due anni successivi al momento della concessione della libertà provvisoria.

Le conclusioni di ProPublica sono sinteticamente espresse in Figura 8.2: in quasi il doppio dei casi rispetto ai Bianchi, Compas classifica gli Afro-Americani come soggetti ad alto rischio di recidiva, senza che abbiano successivamente commesso nuovi reati nei due anni successivi.

Inoltre, in quasi il doppio dei casi rispetto agli Afro-americani, i Bianchi sono classificati come soggetti a basso rischio di recidiva, e hanno commesso nuovi reati nei due anni successivi. Il titolo della figura usa il termine *bias*, che in inglese significa *distorsione* o *discriminazione*.

[12] https://www.propublica.org/

	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%

Figura 8.2 – L'analisi di ProPublica sul bias nella determinazione del rischio di recidiva

Dopo pochi mesi l'azienda produttrice di Compas, la Northpointe, produsse un altro studio, in cui sosteneva che le misure proposte da ProPublica per dimostrare l'esistenza di bias verso gli Afro-Americani non andavano bene, e che, rispetto ad altre misure proposte nello studio, non vi era nessuna discriminazione,

Il Capitolo è dedicato a capire come sia possibile che ProPublica e Norhpointe siano arrivate a queste diverse conclusioni, affrontando in questo modo il tema centrale di questo libro, l'equità dei modelli predittivi che utilizzano tecniche di Machine learning, e quindi il tema dell'etica dei dati e delle macchine; scopriremo che, piuttosto che parlare di una sola equità, dobbiamo parlare di diverse equità, in alcuni casi inconciliabili tra di loro. Ciò ci permetterà di comprendere che l'equità è sì un tema etico, ma che non esiste una sola equità assoluta e universale, piuttosto *diverse forme di equità tra cui dobbiamo volta a volta scegliere a seconda delle nostre convinzioni e visione sociale del mondo*.

8.2 La equità o le equità? Il punto di vista di ProPublica

Ricordo che nell metodo di produzione di un modello predittivo, applicando il modello ai dati di test, troviamo veri e falsi positivi e veri e falsi negativi, a seconda della relazione tra rischio iniziale e rischio assegnato, o, equivalentemente, la concessione o meno della libertà provvisoria. La rappresentazione che adotteremo è quella di Figura 8.3 parte destra, sostanzialmente equivalente a quella di Figura 7.3 che mostro nella parte sinistra della figura.

Rischio iniziale basso Si-Libertà provvisoria **Veri negativi**	Rischio iniziale alto No - Libertà provvisoria **Veri positivi**
Rischio iniziale alto Si-Libertà provvisoria **Falsi negativi**	Rischio iniziale basso No - Libertà provvisoria **Falsi positivi**

Figura 7.3

→

	Rischio alto (RA)	Rischio basso (RB)
Non viene concessa la libertà provvisoria	**Veri positivi (VP)**	**Falsi positivi (FP)**
Viene concessa la libertà provvisoria	**Falsi negativi (FN)**	**Veri negativi (VN)**

Figura 8.3 – Nuova rappresentazione degli elementi della Figura 7.3: veri e falsi positivi e negativi riferiti alla concessione o meno della libertà provvisoria, per detenuti giudicati inizialmente ad alto e basso rischio

Cosa significa, a questo punto, che un modello predittivo non è equo? Abbiamo visto che la accuratezza di un modello predittivo, che possiamo chiamare informalmente la *capacità di*

indovinare esattamente, può essere valutata con diverse misure, che abbiamo chiamato inizialmente con lo stesso nome, accuratezza, e con i termini precisione, sensibilità o recall, specificità.

Ciascuna di queste misure osserva un particolare aspetto della accuratezza, ogni volta legato alla relazione che c'è nel campione che prendiamo in considerazione, tra veri positivi, veri negativi, falsi positivi e falsi negativi.

La Figura 8.4 estende la Figura 8.3 con le misure di specificità e sensibilità. La sensibilità è la percentuale dei veri positivi, cui correttamente non viene concessa la libertà, rispetto al totale dei detenuti cui non viene concessa la libertà; per ottenere questo totale dobbiamo aggiungere a denominatore i falsi negativi, che non hanno concessa la libertà perché pur essendo a basso rischio, sono "capitati" in una classe in cui i detenuti ad alto rischio sono prevalenti.

	Rischio alto (RA)	Rischio basso (RB)
Non viene concessa la libertà provvisoria	**Veri positivi (VP)**	**Falsi positivi (FP)**
Viene concessa la libertà provvisoria	**Falsi negativi (FN)**	**Veri negativi (VN)**
Percentuale dei detenuti a cui **non viene** concessa la libertà provvisoria (LP-No) quando il rischio è →	Probabilità che se LP-No allora RA **Percentuale di veri positivi (Sensibilità) VP / (VP + FN)**	
Percentuale dei detenuti a cui **viene** concessa la libertà provvisoria (LP-Sì) quando il rischio è →		Probabilità che se LP-Sì allora RA **Percentuale di veri negativi (Specificità) VN/(FP+VN)**

Figura 8.4 - Sensibilità e specificità nel nuovo quadro

La specificità misura il caso opposto, cioè la percentuale dei detenuti a cui viene correttamente concessa la libertà, rispetto al numero di tutti coloro cui viene concessa la libertà.

In questo quadro, dunque, sensibilità e specificità sono massime (pari ad 1) quando, rispettivamente, *non vengono liberati tutti coloro che sono originariamente ad alto rischio* (Sensibilità), e *vengono liberati tutti coloro che sono originariamente a basso rischio (Specificità).*

Siamo pronti per comprendere dove la denuncia di ProPublica si colloca in questo quadro. Osservate la Figura 8.5, i due riquadri con la cornice blu sono le misure proposte da ProPublica per denunciare la iniquità tra Afro-Americani e Bianchi, espresse nel quadro di misure introdotto poco fa. La percentuale di falsi positivi ci dice quanti detenuti non hanno avuto ingiustamente la libertà provvisoria sul totale, cioè quanti detenuti erano a rischio basso,

eppure, trovandosi in una classe i cui hanno prevalso numericamente i detenuti a rischio alto, hanno la libertà negata.

	Rischio alto (RA)	Rischio basso (RB)
Non viene concessa la libertà provvisoria	**Veri positivi (VP)**	**Falsi positivi (FP)**
Viene concessa la libertà provvisoria	**Falsi negativi (FN)**	**Veri negativi (VN)**
Percentuale dei detenuti a cui **non viene** concessa la libertà provvisoria (LP-No) quando il rischio è →	Probabilità che se LP-No allora RA **Percentuale di veri positivi (Sensibilità) VP / (VP + FN)**	Probabilità che se LP-No allora RB **Percentuale di falsi positivi FP / (FP + VN)**
Percentuale dei detenuti a cui **viene** concessa la libertà provvisoria (LP-Sì) quando il rischio è →	Probabilità che se LP-Sì allora RA **Percentuale di falsi negativi FN / (FN + VP)**	Probabilità che se LP-Sì allora RA **Percentuale di veri negativi (Specificità) VN/(FP+VN)**

Figura 8.5 – Le misure di equità utilizzate da ProPublica

Occorre dire che nello studio di ProPublica si utilizzano dati sulla recidiva raccolti successivamente alla concessione della libertà, e si confrontano con il livello di rischio attribuito al detenuto. Nel nostro modello predittivo si utilizzano i dati di test, e si confrontano il rischio attribuito al detenuto con il rischio dal modello, e quindi con la concessione o meno della libertà. Non c'è sostanziale differenza tra i due quadri.

E adesso calcoliamo le percentuali di falsi positivi e falsi negativi, assumendo come valori iniziali dell'esercizio quelli di Figura 8.6.

Etnia	Numero arresti	Rischio Basso	Rischio Alto	Esito finale rischio	Decisione
AfroAmericani	0	2.400	1.900	Basso	Libertà provvisoria concessa
AfroAmericani	1 o +	1.800	4.100	Alto	Libertà provvisoria non concessa
Bianchi	0	2.700	1.950	Basso	Libertà provvisoria concessa
Bianchi	1 o +	800	2.600	Alto	Libertà provvisoria non concessa

Figura 8.6 – Rischi e decisioni per le diverse tipologie di detenuti nel nostro esempio

Esercizio 8.1 - Costruisci ora l'albero di decisione, calcola il grado di rischio, i veri e falsi positivi e negativi e la decisione finale, e guarda il risultato nella prossima pagina.

Ecco la risposta all'esercizio.

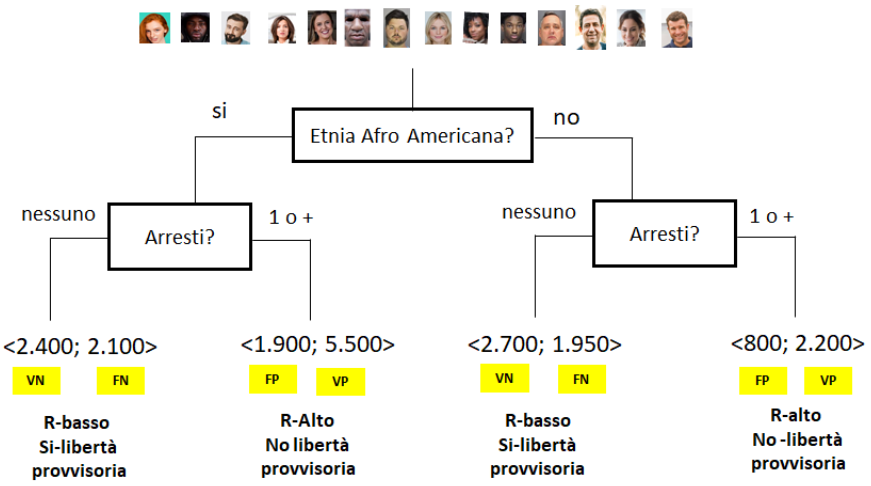


Figura 8.7 - Albero dei casi possibili, relative decisioni e veri e falsi positivi e negativi

Riporto in Figura 8.8 i dati di Figura 8.7 in forma più chiara per calcolare le due misure ProPublica.

Esercizio 8.2 - Ora calcolate voi le percentuali di falsi positivi e negativi per Bianchi e Afro-Americani, confrontandola con i dati del quadro ProPublica.

Etnia	Tipologia Universo	Numerosità
Afro Americani	Veri Positivi	5.500
Afro Americani	Falsi Positivi	1.900
Afro Americani	Veri Negativi	2.400
Afro Americani	Falsi Negativi	2.100
Bianchi	Veri Positivi	2.200
Bianchi	Falsi Positivi	800
Bianchi	Veri Negativi	2.700
Bianchi	Falsi Negativi	1.950

Figura 8.8 - Veri e falsi positivi e negativi nell'esempio

Ecco la risposta in Figura 8.9.

	Rischio alto (RA)	Rischio basso (RB)
Non viene concessa la libertà provvisoria	**AfroAmericani = 5.500** **Bianchi = 2.200**	**AfroAmericani = 1.900** **Bianchi = 800**
Viene concessa la libertà provvisoria	**AfroAmericani = 2.100** **Bianchi = 1.950**	**AfroAmericani = 2.400** **Bianchi = 2.700**
Percentuale dei detenuti a cui **non viene** concessa la libertà provvisoria (LP-No) quando il rischio è →	Probabilità che se LP-No allora RA **Percentuale** **di veri positivi** **(Sensibilità)** **VP / (VP + FN)**	**Percentuale di falsi positivi** Afro Americani = 0,44 Bianchi = 0,23
Percentuale dei detenuti a cui **viene** concessa la libertà provvisoria (LP-Sì) quando il rischio è →	**Percentuale di falsi negativi** Afro Americani = 0,28 Bianchi = 0,47	Probabilità che se LP-Sì allora RA **Percentuale** **di veri negativi** **(Specificità)** **VN/(FP+VN)**

Il nostro studio di caso

	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%

Lo studio di ProPublica

Figura 8.9 – Percentuali simili nei due quadri interpretativi adottati in questo libro e di ProPublica

Dunque, sia pure attraverso modelli diversi, i due casi di studio, quello di ProPublica e quello mostrato qui, arrivano a percentuali simili.

Se noi vediamo l'equità come uguale percentuale di casi tra le due etnie di detenuti scorrettamente valutati ad alto rischio (falsi positivi) e scorrettamente valutati a basso rischio (falsi negativi), allora gli Afro-Americani sono discriminati, e i Bianchi sono favoriti. Insomma, *ci sono più Afro-Americani in percentuale* che non vengono liberati pur essendo a basso rischio, *e più Bianchi* che vengono liberati, pur essendo ad alto rischio.

Trovo tutto ciò ingiusto per gli Afro-Americani! Mi chiedo: perché questa situazione di ingiustizia verso gli Afro-Americani? C'è qualcosa che mi sfugge: la tecnica degli alberi di decisione mi sembra neutrale tra Afro-Americani e i Bianchi; non sembra che faccia discriminazioni. Dove sta l'inghippo?

Per capire tutti i motivi per cui i modelli predittivi basati su Machine learning creino iniquità, occorre attendere i Capitoli 9 e 10. Ma una risposta te la voglio dare subito.

Guardiamo la Figura 8.10.

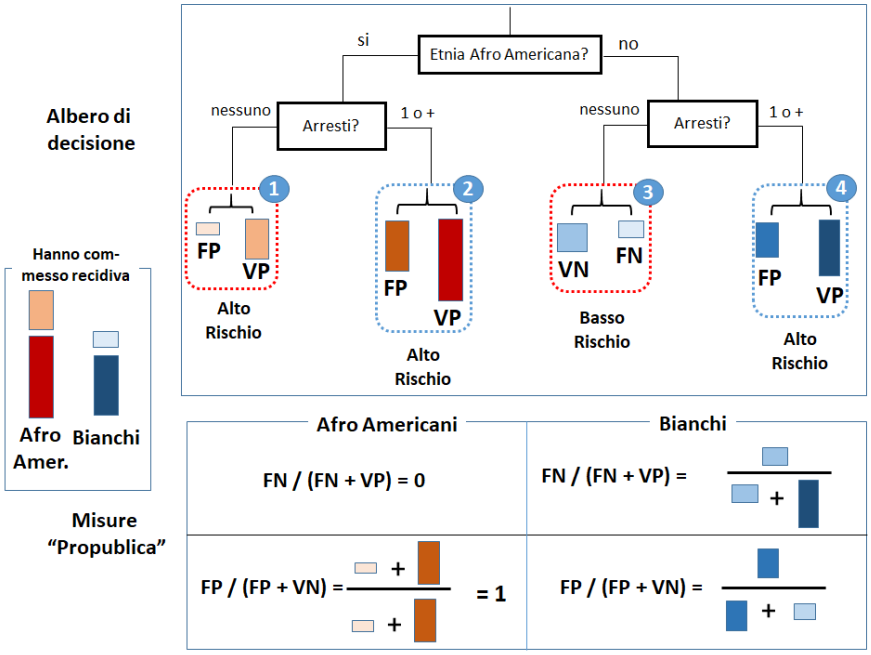


Figura 8.10 – Tassi di recidiva nelle persone presenti nei dati di addestramento, albero di decisione e misure "ProPublica" rappresentati in forma visuale anziché con numeri.

Anzitutto, ti devo dare una informazione importante; nel periodo storico a cui i dati fanno riferimento, *il tasso di recidiva era più alto nel caso degli Afro-Americani rispetto ai Bianchi*, insomma gli Afro-Americani *tendevano a essere arrestati per aver commesso reati più dei Bianchi*. Ho mostrato questo nella parte sinistra della figura, in cui i rettangolini di colore base rosso sono i gruppi di Afro-Americani con recidiva nell'albero di decisione, e i rettangolini di colore base blu sono i gruppi di Bianchi.

Come si vede, i maggiori tassi di recidiva portano *per entrambe le foglie* associate agli Afro-Americani (foglia 1 e foglia 2) a far prevalere gli elementi ad alto rischio, mentre per i Bianchi in un caso prevale il basso rischio (foglia 3) e nell'altro l'alto rischio (foglia 4).

Ciò permette di capire perché gli Afro-Americani avessero una percentuale di falsi negativi (cioè di elementi che erano ad alto rischio e che venivano liberati) più bassa dei Bianchi (nell'esempio addirittura nessuno!); e avessero una percentuale di falsi positivi (cioè di elementi a basso rischio che vengono classificati ad alto rischio e quindi non vengono liberati) più alta.

Va bene, capisco tutto, ma il problema rimane: oggi, non allora, gli Afro-Americani sono discriminati!

E’ vero. Come dicevo, devi avere pazienza, riprenderemo il discorso prima della fine del libro, quand avremo visto tutte le possibili cause di discriminazione e un po' di rimedi.

Notate che i due indicatori scelti da ProPublica possono essere visti come due misure di *accuratezza*, ulteriori rispetto a quelle viste nel capitolo precedente; notate anche che i due restanti riquadri in basso rappresentano la sensitività e la specificità.

Insomma, l’esempio ProPublica ci dice che quando noi vogliamo verificare se è commessa qualche ingiustizia *rispetto a una specifica parte della popolazione* su cui è applicato il modello predittivo, non dobbiamo applicare il modello predittivo alla popolazione indistinta, come in Figura 8.11, ma dovremo, piuttosto, calcolare i valori di accuratezza per la specifica parte della popolazione (nel nostro esempio, gli Afro-Americani) e per la parte restante (nel nostro esempio, i Bianchi), e confrontare i valori ottenuti per le due classi; queste due classi, in Figura 8.12, possono essere, ad esempio, le donne e gli uomini, i Bianchi e gli Afro-Americani, o altre categorie di persone per le quali sono definite specifiche protezioni sociali.

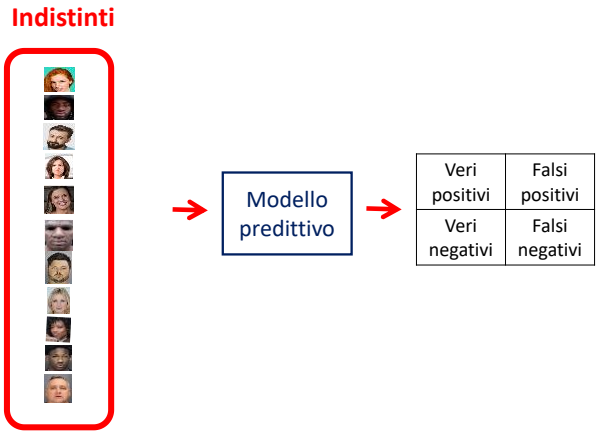


Figura 8.11 – Calcolo delle misure di accuratezza indipendentemente da specifici gruppi e scopi

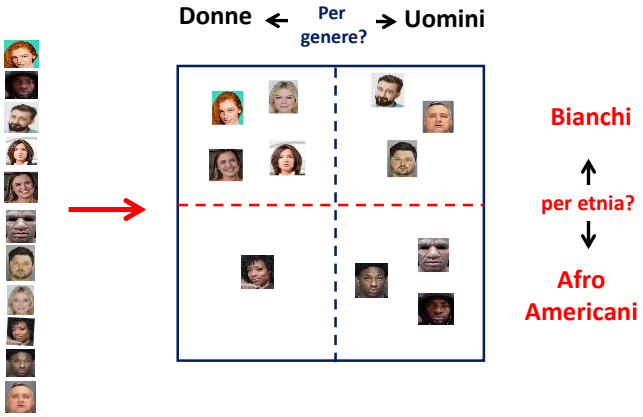


Figura 8.12 – Composizione della popolazione sulla base di due diverse caratteristiche

In Figura 8.13 vediamo il caso più generale, in cui possiamo avere la necessità di misurare la equità per differenti gruppi sociali e per differenti scopi dei modelli predittivi.

Figura 8.13 – Calcolo della equità nei due gruppi per differenti scopi

8.3 La reazione di Northpointe alla denuncia di ProPublica

Nello stesso anno, il 2016, in cui ProPublica denunciava la discriminazione verso gli Afro-Americani, l'azienda che produceva Compas, la Northpointe, commissionò uno studio i cui risultati erano in totale contrasto con la denuncia di ProPublica: non c'è nessuna discriminazione verso gli Afro-Americani, ProPublica sbaglia.

L'analisi di Northpointe focalizzava l'attenzione su indicatori diversi rispetto a quelli di ProPublica, vedi Figura 8.14. Il ragionamento era questo: ProPublica cerca di capire come sono distribuiti tra Afro-Americani e Bianchi i detenuti falsi positivi e negativi che rispettivamente hanno o non hanno avuto concessa la libertà provvisoria; proviamo invece a partire dal livello di rischio, e concentriamoci non sui falsi positivi e negativi, *ma sui veri positivi e veri negativi*, sulla popolazione, cioè, per la quale *correttamente si decide la concessione o non concessione della libertà*. Nelle cornici rosse vengono riportati direttamente i dati dello studio Northpointe, non faremo un nostro esempio.

Propublica Northpointe

	Rischio alto (RA)	Rischio basso (RB)	Percentuale dei detenuti che con **rischio alto** hanno come esito LP →	Percentuale dei detenuti che con **rischio basso** hanno come esito LP →
Non viene concessa la libertà provvisoria	**Veri positivi (VP)**	**Falsi positivi (FP)**	**Capacità predittiva dei positivi (Precisione) VP/(VP + FP)** Afro Americani = 0,73 Bianchi = 0,74	
Viene concessa la libertà provvisoria	**Falsi negativi (FN)**	**Veri negativi (VN)**		**Capacità predittiva dei negativi VN/(VN + FN)** Afro Americani = 0,58 Bianchi = 0,53
Percentuale dei detenuti a cui **non viene** concessa la libertà provvisoria (LP-No) quando il rischio è →		**Percentuale di falsi positivi FP/(FP + VN)** Afro Americani = 0,44 Bianchi = 0,23		
Percentuale dei detenuti a cui **viene** concessa la libertà provvisoria (LP-Si) quando il rischio è →	**Percentuale di falsi negativi FN/(FN + VP)** Afro Americani = 0,28 Bianchi = 0,47)	**Percentuale di veri negativi VN/(FP+VN)**		

Figura 8.14 – Gli indicatori di Northpointe

Gli indicatori Northpointe sono chiamati *capacità predittiva dei positivi e dei negativi*, perché esprimono la capacità del modello predittivo di indovinare i detenuti a cui è giusto non concedere e concedere la libertà provvisoria. Northpointe dice in sostanza: è a queste percentuali che bisogna guardare, sono questi i risultati di interesse per il giudice! E, aggiunge: rispetto a questi, Compas funziona, non è discriminatorio.

Non ci capisco più niente! Prima arriva ProPublica e dice: gli Afro-Americani sono discriminati, poi arriva Northpointe e dice: non è vero che gli Afro-Americani sono discriminati, mi chi ha ragione?

Capisco il tuo sconcerto! Ti comincio a rispondere nella prossima sezione.

8.4 Le tante definizioni di Equità

Guarda in Figura 8.15 come sono cresciute nel tempo le definizioni di equità, tutte diverse l'una dall'altra, negli anni dal 2012 al 2017: come minimo una crescita di un fattore uno a 20!

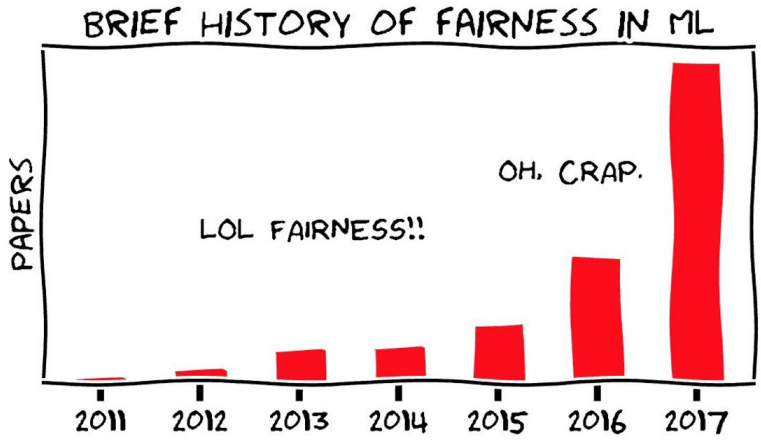


Figura 8.15 - La crescita delle definizione di Equità nella letteratura, dal 2011 al 2017

Quando ho visto questa figura, mi sono detto: cerchiamo di capire bene come sono organizzate queste diverse definizioni, a quali principi si ispirano. Guidato soprattutto dal lavoro [13], ho prodotto la classificazione di Figura 8.16, in cui ho evidenziato in rosso i concetti trattati in questo libro.

[13] S. Verma et al. Fairness Definitions Explained - 2018 ACM/IEEE International Workshop on Software Fairness.

Tipi di equità (fairness) basati sul a. valore positivo/negativo predetto per gli elementi e b. sul valore degli elementi nella realtà	Definizione
Misure di base 1. VP - Veri Positivi 2. VN - Veri Negativi 3. FP - Falsi Positivi 4. FN - Falsi Negativi	1. Elementi (della popolazione) che sono predetti positivi e che sono positivi nella realtà 2. Elementi che sono predetti negativi e che sono negativi nella realtà 3. Elementi che sono predetti positivi e che sono negativi nella realtà 4. Elementi che sono predetti negativi e che sono positivi nella realtà
Misure statistiche di accuratezza 1. **VPP** – Valore predittivo dei positivi VP/(VP+FP) **Precisione → C7** 2. CSN – Capacità di scoprire I negativi FP/(VP+FP) 3. ISN - Incapacità di scoprire I negativi FN/(VN+FN) 4. VPN - Valore predittivo dei negativi VN/(VN+FN) 5. **PVP** – Perc. di veri positivi VP/(VP+FN) **Recall/Sensitività → C7** 6. PFP - Percentuale di falsi positivi FP/(FP+VN) 7. **PFN** - Percentuale di falsi negativi FN/(FN+VP) **Specificità → C7** 8. PVN - Percentuale di veri negativi VN/(FP+VN)	1. Probabilità che un elemento predetto positivo sia positivo nella realtà 2. Probabilità di un negativo di essere predetto scorrettamente come positivo 3. Probabilità di un caso positivo di essere negativo nella realtà 4. Probabilità che un elemento predetto negativo sia negativo nella realtà 5. Probabilità di un vero positivo di essere identificato come tale 6. Probabilità di attribuire un valore positivo ad un elemento negativo 7. Probabilità di attribuire un valore negativo ad un elemento positivo 8. Probabilità di un vero negativo di essere identificato come tale
Basate sul valore predetto per gli elementi per varie distribuzioni demografiche degli elementi 1. Equità di gruppo / Parità statistica / **Parità demografica → Capitolo 8**	1. Gli elementi nelle popolazioni protette e non protette hanno uguale probabilità di essere assegnati alla classe dei positivi – **Obiettivo della Società**
Basate sul confronto tra gruppi protetti e non protetti sui valori predetti e valori nella realtà 1. **Parità predittiva → Capitolo 8** 2. **Uguale errore sui falsi positivi → Capitolo 8** 3. **Uguale errore sui falsi negativi → Capitolo 8** 4. **Uguali quote di positive → Capitolo 7**	1. I gruppi protetti e I gruppi non protetti hanno uguale **VPP Precisione - Obiettivo del Decisore & Northpointe** 2. I gruppi protetti e I gruppi non protetti hanno uguale **PFP - Obiettivo del Detenuto & Propublica** 3. I gruppi protetti e I gruppi non protetti hanno uguale **PFN - Obiettivo del Detenuto & Propublica** 4. I gruppi protetti e I gruppi non protetti hanno uguale **PVP Sensitività**
Basate su ragionamento causale (Perchè?) 1. **Equità ottenuta attraverso la non consapevolezza → Capitolo 10** 2. Equità ottenuta attraverso la consapevolezza	1. Nessun attributo sensibile è utilizzato nel processo predittivo 2. Elementi simili secondo una funzione di distanza devono avere valori simili nella predizione

Figura 8.16 – Definizioni dei principali tipi di equità (in rosso le tipologie citate nel libro e il Capitolo dove si trova la tipologia, C significa Capitolo)

Tra le definizioni di Figura 8.16 e finora non commentate è importante la *Parità demografica*, che possiamo spiegare con un esempio. Supponiamo che ad un concorso si presentino 900 uomini e 100 donne, e che ci siano 10 vincitori. Siccome è noto che le donne rappresentano il 50% della popolazione complessiva, supponendo che per decidere la grauatoria sia adottato un modello classificatorio basato su esempi del passato, la parità demografica è raggiunta quando il modello classificatorio attribuisce i posti equamente agli uomini (5 vincitori) e alle donne (5 vincitrici).

Tutte le definizioni di Figura 8.16 hanno in comune l'idea di individuare forme di equità che riguardano gruppi sociali uno dei quali è visto, a seguito di leggi o per una diffusa sensibilità sociale, come categoria da proteggere. Chiediamoci ora: quale definizione adottare in un determinato contesto?

Torniamo alla disputa tra ProPublica e Northpointe, e chiediamoci: chi ha ragione? E' la stessa domanda che si fece un giornalista del Washington Post, che, studiate le due analisi prodotte da ProPublica e Northpointe disse: dipende! Certo, dipende da quale definizione si adotta di equità, non è possibile che due definizioni diverse di equità portino agli stessi risultati, è evidente dal nostro esempio, e si potrebbe dimostrare in modo formale.

Nel caso ProPublica Northpointe: o vale la percentuale di falsi positivi e la percentuale di falsi negativi, oppure valgono le due capacità predittive dei positivi e dei negativi. Non c'è via di mezzo!

Ma allora, è solo questione di punti di vista? Insomma, dobbiamo tirare una monetina e vedere se esce testa o croce?

Assolutamente no! Se pensi per un'attimo alla definizione di equità di ProPublica, l'idea è che *bisogna proteggere tutte le popolazioni appartenenti a etnie diverse dall'avere negata la libertà provvisoria perché sono classificate ad alto rischio, ma in realtà non lo sono;* questa definizione è, insomma, dalla parte dei detenuti, li vuole cautelare sul fatto che non saranno discriminati dai dati raccolti nel passato, si potrebbe dire "le colpe dei padri non ricadano sui figli"!

Mentre invece la definizione di Northponte, uso un termine irrituale, strizza l'occhio al giudice, sembra dirgli: il tuo interesse è cercare di prevedere il futuro nella maniera più esatta possibile, cercare di prevedere quali detenuti potranno commettere nuovamente un reato, e quali non lo commetteranno; la previsione è fatta su quanto accaduto nel passato, è basata sui *precedenti*. Insomma, Northpointe sembra assumere il punto di vista di *chi deve giudicare*.

Ma anche dal punto di vista di chi deve giudicare, se posso esprimere una opinione anche non facendo come professione il giudice, cercherei di conoscere non solo le percentuali sulla capacità predittiva (equità Northpointe), ma anche quelle sui falsi positivi e negativi (equità ProPublica), e poi deciderei, magari applicando, a parità di altri elementi, il principio: in dubio pro reo.

Si, tutto chiaro, tutto bene, ma con questo capitolo del libro mi hai completamente stravolto tante certezze che avevo! Io ho sempre creduto che ci fosse un solo tipo di equità…

Non è così, e non ci posso far niente: la novità è che *ne esistono tante*, così come sono tante e diverse le nostre aspettative. La questione è simile a quanto accade con il concetto di valore, a cui sarà dedicato un libro della Enciclopedia, e che ora brevemente accenno.

Adam Smith introdusse il concetto di valore d'uso negli scambi tra esseri umani; gli scambi possono avvenire con il baratto, in cui tra due soggetti ci si scambia beni in natura, oppure con un acquisto, in cui si scambia un bene (due chili di arance) o un servizio (ad es. una prenotazione a uno spettacolo) con una quantità di denaro (3 euro).

Il valore d'uso che una persona associa a un bene o un servizio può essere valutato come una relazione tra:

- Il beneficio che traggo dal bene o servizio
- Il sacrificio che devo compiere.

Per esempio, quando io compro due chili di arance per 3 euro, il beneficio è che le posso mangiare o spremere per fare una aranciata, il sacrificio è in parte la somma che devo spendere, in parte il tempo che devo perdere per andare al mercato o al supermercato.

In realtà le cose non sono mai così semplici: a me per esempio non piacciono le arance Navel, e piacciono le Tarocco rosse a buccia fine, e so che ho più probabilità di trovarle al

mercato che al supermercato, ma so anche che devo perdere più tempo al mercato che al supermercato; alla fine soppeso mentalmente benefici e sacrifici, e decido (magari compro a qualche frazione di euro in più la spremuta d'arancia, così risparmio il tempo per mettere su la spremiagrumi, e soprattutto per pulirla).

Infine, e non ultima, i 3 euro sono *una inezia* per una persona agiata, possono essere una cifra rispettabile e talvolta irraggiungibile per un migrante o una persona anziana con una pensione minima.

Insomma, *siamo tutti diversi e abbiamo diverse sensibilità e diversi tenori di vita rispetto ai beni che acquistiamo e ai servizi che fruiamo.*

Anche riguardo alle diverse definizioni di equità ci sono interessi diversi: è interesse del *detenuto* non essere considerato falsamente ad alto rischio, e quindi non ottenere la libertà, è interesse del *giudice* massimizzare la probabilità di decidere bene, prevedendo ciò che potrà accadere in futuro. E' interesse della *società*, vedi ancora Figura 8.16, che le decisioni abbiano un impatto uguale tra le sue diverse componenti.

Attenzione però: *il giudice deve essere consapevole che le previsioni sul futuro formulate da Compas o da altro modello predittivo sono affette da errore, e, soprattutto, sono basate su dati del passato*.

Facciamo come abbiamo sempre fatto, sembra dire Compas: ma questo non è l'unico modo di procedere. Ricordate la figura sul mondo come dovrebbe essere, come è, come è visto nella sua rappresentazione digitale? La riproduco per comodità qui di seguito, vedi Figura 8.17.

Quale mondo deve avere come riferimento il giudice che deve decidere se concedere o meno la libertà provvisoria a un detenuto che ha presentato domanda, con cui ha avuto un colloquio da cui ha tratto una certa impressione, di cui ha letto la vicenda umana? Un giudice che abbia anche a disposizione uno strumento tipo Compas, ma che vuole decidere per conto proprio? Può fare riferimento a tre mondi, dal basso in alto.

1. Il *mondo di Compas*, strumento che acquisisce un certo numero di caratteristiche del detenuto (data di nascita, quanti arresti ha subito nel passato, se è uomo o donna, ecc. ecc.), le stesse caratteristiche usate nel passato per profilare come ad alto rischio o a basso rischio i detenuti.

2. il *mondo reale*, fatto del detenuto in carne ed ossa, fatto delle cose che dice, delle sue relazioni familiari, della sua personalità come si è prodotta nelle esperienze di vita.

3. il *mondo ideale*, come dovrebbe essere secondo le nostre visioni, e come potrebbe essere se non lo subiamo solo passivamente, ma *interveniamo* per cambiarlo, nelle nostre azioni giornaliere, nella vita di ogni giorno o nel nostro lavoro, nel nostro impegno sociale e politico.

C'è una tensione tra questi tre mondi, e c'è una tensione tra le diverse equità e i soggetti coinvolti. Tutte le volte che viene usato un modello predittivo, c'è un soggetto che lo usa

per prendere una decisione, e un secondo soggetto, o una popolazione di soggetti, che subisce questa decisione.

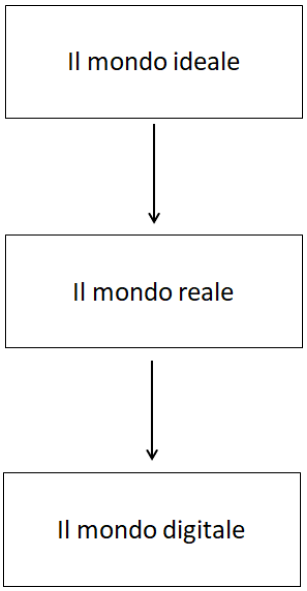


Figura 8.17 – Quanti mondi ci sono in una decisione….

Se chiediamo un prestito a una banca, il direttore della banca deve decidere se erogarlo o meno; in questo caso chi prende la decisione è il direttore della banca, chi la subisce, chi è coinvolto direttamente nella decisione, che, a seconda che sia positiva o meno, potrà permettere di iscriversi a una università o acquistare una casa o solo affittare una casa, siamo noi. In questo caso io ho diritto a essere trattato nello stesso modo di tutte le altre persone che si trovino nella mia stessa situazione, senza favoritismi verso altri, e senza favoritismi verso di me. E senza discriminazione dovuta a genere, etnia, ecc.

Se a chiedere un prestito sono un gruppo di cittadini che per un terremoto hanno visto la loro casa distrutta, il decisore è sempre il direttore della banca, i soggetti coinvolti dalla decisione sono la popolazione dei cittadini che ha chiesto il prestito. La popolazione dei cittadini ha diritto di essere trattata allo stesso modo di altre popolazioni, e anche i singoli cittadini hanno il diritto di essere trattati allo stesso modo, senza favoritismi.

Mah, non sono d'accordo che uno che ha perso tutto debba essere trattato allo stesso modo di uno che ha un alto tenore di vita!

Certo, si dovrebbe avere un occhio di riguardo, è quella che viene chiamata *equità riparativa* di cui parlerò nel Capitolo 10.

Conclusioni

La vita è fatta di tensioni tra interessi diversi e diverse visioni del mondo, possiamo sintetizzare così tutto quanto abbiamo visto in questo capitolo. Non esiste una sola definizione di equità, ne esistono tante, alcune in conflitto tra di loro. Ognuna risponde ad un principio, a un modo per esprimere l'assenza di discriminazioni.

Quando usiamo un modello predittivo, prima di considerarlo per prendere una decisione, cerchiamo anzitutto di capire che ruolo svolgano e che valori abbiano le diverse definizioni di equità a nostra disposizione, e chi siano i soggetti interessati all'una o all'altra.

Cerchiamo di capire in modo trasparente come il modello si comporti rispetto alle diverse definizioni di equità, e le eventuali discriminazioni a quali soggetti si rivolgano. E proviamo a mitigare le discriminazioni, con le regole che vedremo nel Capitolo 10.

Capitolo 9. Questioni etiche che coinvolgono il modello, la tecnica, i dati e gli esseri umani

Quanto visto finora assume che il *mondo dei dati digitali* osservato dal modello predittivo sia aderente al mondo reale, così come si è manifestato nel passato e come si manifesta nel presente. Ma, fin dal primo capitolo del primo libro di questa Enciclopedia, abbiamo visto che così non è, i dati digitali sono una rappresentazione del mondo imprecisa, distorta, parziale.

Quei dati digitali del passato su cui si basano le previsioni del rischio sono stati prodotti da esseri umani, attraverso le loro decisioni; il modello predittivo che è stato scelto per prevedere il futuro rischio, è a sua volta affetto da errori, come abbiamo visto in maniera evidente in tutti gli esempi del Capitolo 7 e del Capitolo 8; quell'albero di decisione ha generato valutazioni di rischio e decisioni di libertà provvisoria distorte, che generano falsi positivi e falsi negativi.

In questo Capitolo affrontiamo in maniera più sistematica le questioni etiche coinvolte nei modelli predittivi generati da tecniche di Machine learning. Consideriamo dapprima la questione generale dell'impatto sugli esseri umani delle predizioni effettuate dai modelli. Poi passiamo a considerare separatamente i blocchi della Figura 9.1 discutendoli nei seguenti aspetti:

1. la rappresentazione del mondo attraverso i *dati;*
2. le attività di modellazione, predizione e decisione insieme;
3. il feedback sui dati.

A questo punto discuteremo quali azioni di regolazione stia predisponendo l'Unione Europea in tema di uso di modelli predittivi. Infine tratteremo una questione filosofica fondamentale: i limiti degli modelli predittivi basati sul Machine learning, e la necessità di guardare oltre, e investigare il *perché* delle cose.

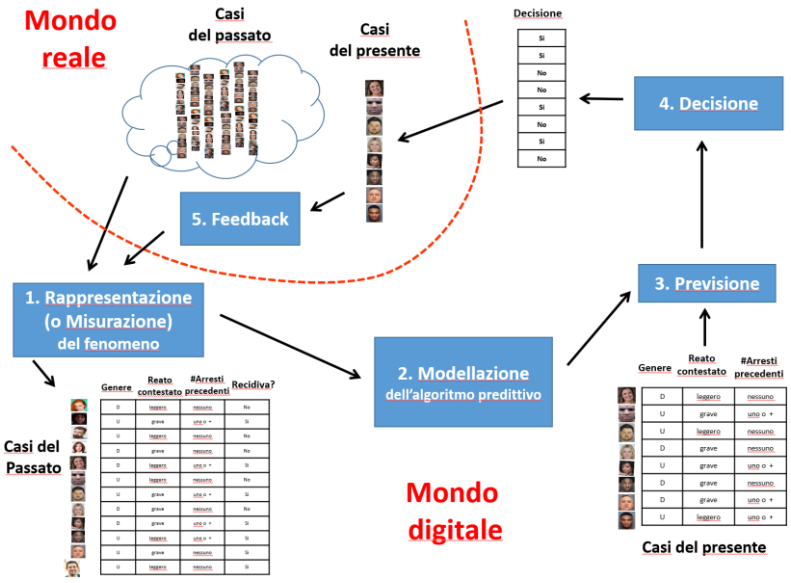


Figura 9.1 - I quattro stadi del Machine learning etico

9.1 I modelli predittivi riguardano gli esseri umani

L'aspetto più delicato dei modelli classificatori e predittivi sta nel fatto che riguardano *quasi sempre gli esseri umani*; i modelli influiscono a volte drammaticamente sulla vita delle persone. In Figura 9.2 vediamo due esempi di tale influenza.

A sinistra c'è il caso che conosciamo bene; se il modello predittivo decide che una persona è ad alto rischio, il giudice può decidere per un supplemento di istruttoria, ma questa decisione è legata a tanti fattori, quali:

- il carico di lavoro (ho tanto da fare, mi fido del modello predittivo),
- la sensibilità etica (d'accordo, ho tanto da fare, ma devo trovare il tempo per approfondire, in fondo il modello predittivo ha deciso in una situazione molto marginale..),
- la conoscenza sulla precisione con cui il modello predittivo ha deciso,
- la curiosità intellettuale del giudice,
- altri fattori imponderabili,

per una decisione che influisce sulla vita di una persona.

Il secondo caso in Figura 9.2 riguarda l'uso di modelli classificatori per il riconoscimento facciale a fini di identificazione delle persone di etnia uigura in Cina; in questo caso i modelli predittivi sono utilizzati per identificare una etnia, e per esercitare sulle persone azioni di isolamento e repressione.

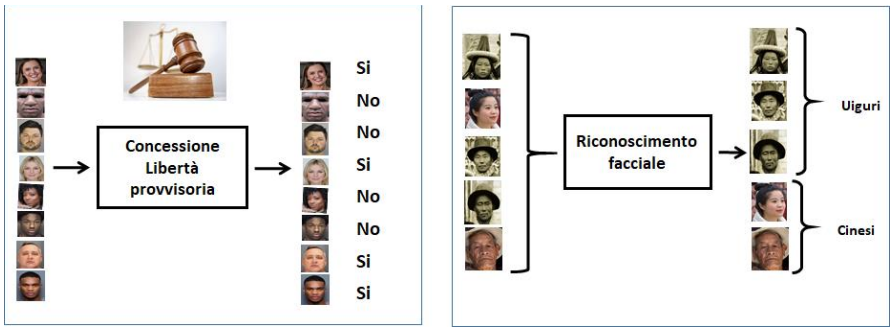


Figura 9.2 - I modelli predittivi riguardano le persone…..

La Figura 9.3 fa riferimento a un esempio già discusso nel Libro 1 della Enciclopedia, e riguarda una iniziativa della città di Boston che aveva lo scopo di individuare con una app su telefono mobile le buche nelle strade della città.

Per rilevare automaticamente le buche, la municipalità di Boston rilasciò una applicazione per telefoni mobili dotati di accelerometro, *ancora poco diffusi allora,* in grado di rilevare improvvisi sobbalzi nelle automobili i cui penumatici incontravano la buca; i sobbalzi generavano un improvviso cambiamento della velocità e accelerazione della automobile.

La sperimentazione portò a risultati distorti, che rilevavano molte buche nei quartieri più agiati, e pochi nelle periferie popolari. Secondo te perché?

Mah, posso immaginare che ciò derivi dal fatto che i telefoni mobili dotati di accelerometro all'epoca avevano un costo che poteva essere sopportato solo da ceti agiati, e quindi erano diffusi in modo ineguale nella città. Ho indovinato?

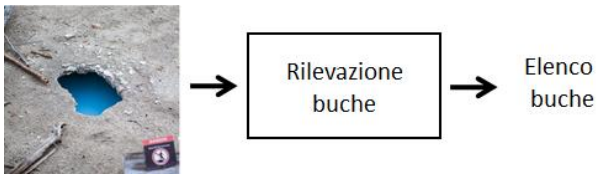


Figura 9.3 - … anche quando non sembra

Sì, hai indovinato! Come vedi, questo caso apparentemente non riguarda le persone, ma, approfondendo, riflette la distribuzione della ricchezza in una popolazione urbana; potemmo dire che c'è un bias di reddito….

9.2 La rappresentazione del mondo nei dati digitali

I dati digitali sono una rappresentazione del mondo. Approfondiamone il ruolo nel nostro contesto.

1. I dati sono scelti dagli esseri umani e riflettono la loro cultura

Riprendiamo il caso Compas e mostriamo in Figura 9.4 le due coordinate che hanno ispirato le scelte fondamentali fatte a suo tempo dalle persone che hanno raccolto i dati: le persone e le caratteristiche attraverso cui rappresentarle.

Chi ha raccolto i dati, ha avuto una grande discrezionalità nello scegliere i detenuti e le caratteristiche usate nelle fonti su cui far apprendere la tecnica predittiva.

Mah, non vedo il problema, basta raccogliere tutte le caratteristiche per tutte le persone cui è stata concessa la libertà provvisoria….

Eh, non è così semplice! Raccogliere dati è una attività faticosa e costosa, in genere viene svolta su un campione di soggetti limitato, il campione: chi ci dice che nella scelta del campione non sia stata fatta una selezione sulle persone discriminatoria?

Non è una questione di sospetti: il problema è che se non sappiamo come è stato scelto il campione dei detenuti, come facciamo a sapere se il campione è sufficientemente rappresentativo dell'universo, o è distorto?

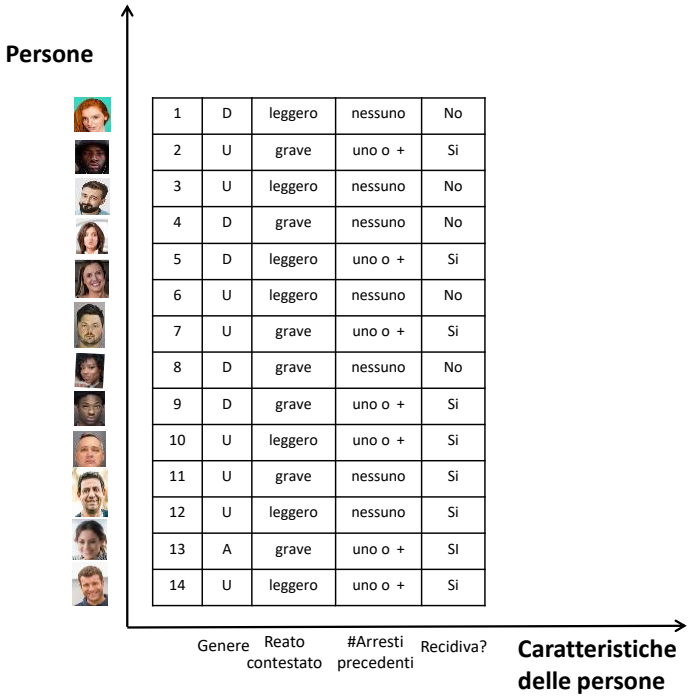


Figura 9.4 – Cosa vogliamo davvero rappresentare con i dati?

2. La variabile obiettivo è la più difficile da definire

Riconsideriamo la Figura 9.1. Quale è, secondo te, lo scopo ultimo del modello predittivo?

Mah… non capisco perché me lo chiedi, dovrebbe essere ormai chiaro: raccogliere un insieme di informazioni su detenuti che a suo tempo hanno ottenuto la libertà provvisoria, per capire se successivamente hanno commesso reati, e costruire così un modello predittivo che calcola il rischio di commettere recidiva! Perché adesso poni di nuovo questa domanda?

Perché, se ci pensiamo, non è evidente cosa significhi e come si misuri il *commettere recidiva*. Quello che si *vorrebbe veramente sapere, per capire se liberare o meno,* è *quanto sia pericoloso il detenuto per la società*; ma questo lo si può sapere *alla fine del processo penale*, fino ad allora la legge dice che deve considerarsi innocente. Ci sono molti casi di persone che vengono arrestate, e successivamente vengono prosciolte, alla fine delle indagini preliminari o al termine del processo penale.

Scusa, ma se io aspetto la fine del processo, non potrò mai produrre un modello predittivo….

Certo, ma è questo il punto, determinare il rischio sulla base del solo aver commesso recidiva, non è *completamente corretto*, è una specie di macchia sulla persona che può essere fissata o lavata solo alla fine del processo, non prima! E, d'altra parte, noi sappiamo che *il modello predittivo sbaglia*: c'è una percentuale di soggetti che sono a basso rischio, a cui però viene assegnato dal modello predittivo un alto rischio.

Così come all'epoca del Covid chi ha fatto un test rapido sapeva che c'era il 30% di falsi negativi, il giudice, come minimo, quando decide deve sapere che c'è una data percentuale di falsi positivi a basso rischio, ma a cui viene assegnato alto rischio.

Sì, effettivamente chi vende un modello predittivo dovrebbe dichiarare la precisione…..

Ti dirò di più: dovrebbe essere *obbligato* da contratto ad usare il modello sui dati di test per verificare il livello delle diverse misure di accuratezza.

Ma scusa, non lo fa già? Chi sviluppa un vaccino, lo deve fare per avere la autorizzazione dagli enti certificatori….

Esatto: per i vaccini esiste una procedura di certificazione, per tutte le altre utilizzazioni no! Vedremo tra poco cosa sta facendo la Unione Europea su questi aspetti.

3. Separa le responsabilità nel suddividere i dati di addestramento e i dati di test

Parto un po' da lontano. Quando si sviluppa un *programma software*, occorre sottoporlo a test, per verificare se fa quello che deve fare. Ad esempio, un programma che sceglie il maggiore tra due numeri interi, deve essere verificato, per esempio, con i dati <3,2>, <2.3>, <3,3>, insomma con diverse tra le combinazioni di casi possibili.

Nei modelli predittivi, il test consiste nel separare i dati storici in due insiemi, dati di addestramento e dati di test, vedi nuovamente Figura 9.5. Dopo aver costruito sui dati di addestramento il modello predittivo, occorre eseguire il modello sui dati di test, e verificare che si riferiscano allo stesso universo di persone, con percentuali delle diverse tipologie simili.

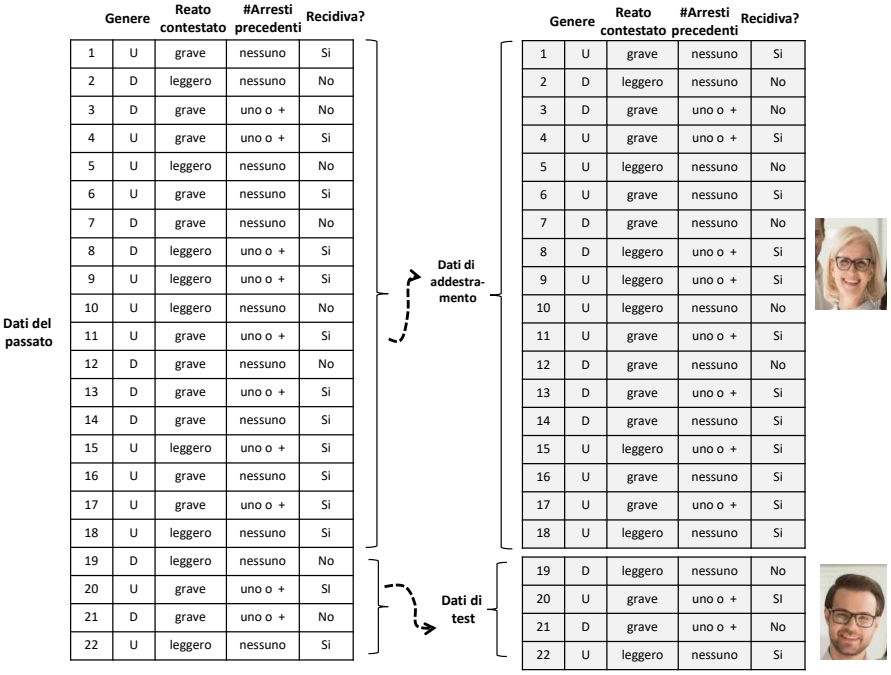


Figura 9.5 – I dati sul passato devono essere ripartiti tra dati di addestramento e dati di test da persone diverse

Per ottenere ciò, i due insiemi di dati devono *scelti e gestiti da persone diverse*; sa nel software che nei modelli predittivi, ciò si ottiene anzitutto applicando il principio della "separation of concerns", la separazione degli interessi, per cui chi sceglie i dati di test è persona diversa rispetto a chi opera sui dati di addestramento. Ciò che si vuole evitare è che vi sia una commistione di interessi, per cui chi opera sui dati di addestramento scelga "maliziosamente" dati di test addomesticati.

Questo è particolarmente importante quando si usino alberi di decisione, perché gli alberi tendono a funzionare bene su dati di addestramento, meno bene su dati aventi caratteristiche diverse e nuove. Detto in un altro modo, il rischio, con gli alberi di decisione, è che si adattino troppo bene ai dati di addestramento (questo comportamento è chiamato di *overfitting*) e non funzionino bene rispetto a dati nuovi. E ciò deve emergere dal processo di test.

3. I dati hanno una storia

I dati hanno una storia, sono raccolti da fonti diverse e fanno riferimento a periodi temporali in genere diversi. Guardiamo la Figura 9.6; i dati usati dal metodo di apprendimento sono stati raccolti in questo caso in quattro anni diversi, nel 2014, nel 2015, nel 2018 e nel 2020.

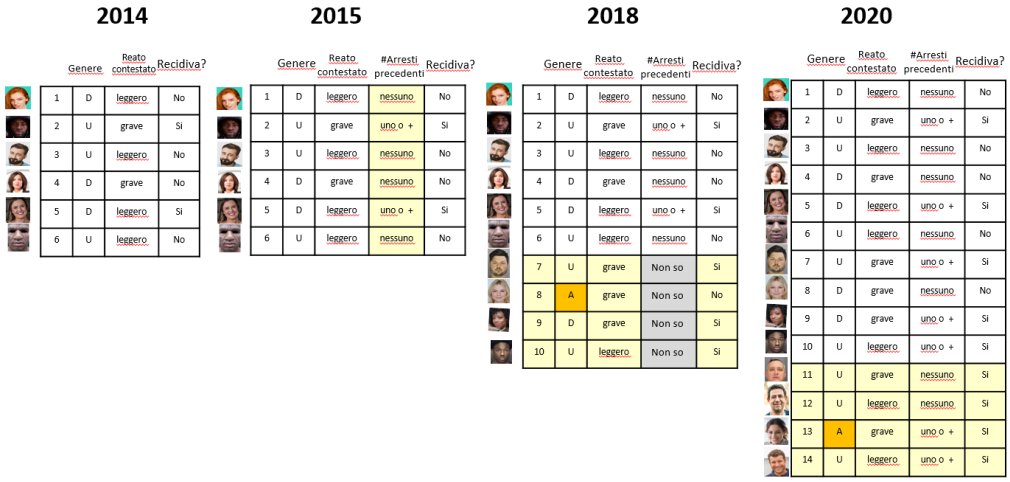


Figura 9.6 – I dati hanno una storia

Nel 2014 sono stati raccolti i dati di cinque detenuti, riguardanti il genere e il reato contestato; assumiamo che l'informazione se abbiano commesso o meno recidiva venga aggiunta assumendo un periodo temporale di tre anni, acquisendola dal sistema penitenziario tramite l'evento di ri-arresto.

Come sono state raccolti durante il 2014i valori di genere e reato? Quale è la fonte? Ci sono varie possibilità: la prima è chiedere al detenuto di compilare un questionario, la seconda è acquisirle ad esempio dalle anagrafi.

Nel 2015 è acquisita la informazione sul numero di arresti precedenti. Questo dato, a differenza degli altri, è un dato *calcolato*: se siano 0, uno, o cinque, lo si può sapere solo

accedendo alle basi di dati che rappresentano gli eventi di arresto e sommando l'insieme degli eventi.

Nel 2018 vengono aggiunti nuovi detenuti; qui accade per la prima volta che il genere possa assumere un *valore ulteriore rispetto a uomo e donna*. In questo caso il valore "altro" porta a mettere in questa categoria persone che hanno vissuto, e intendano rendere pubblica, una esperienza di maturazione della propria identità di genere. Per rispettare questa scelta, viene creato il terzo valore "altro". Questo provoca una modifica dell'insieme dei valori per la caratteristica *genere* nel modello predittivo, modifica che non è semplice gestire, perché tutti i dati usati nel passato non valgono più, e si deve perciò produrre un nuovo modello predittivo.

Un altro problema cui dobbiamo fare attenzione sta nel fatto che negli anni si modifica la popolazione dei detenuti liberati per cui raccogliamo i dati; questa modifica deve essere fatta con attenzione: ad esempio non può cambiare il periodo di osservazione sugli eventi di riarresto.

Un ultimo problema sta nei valori *non so* associati al numero di arresti. E' chiaro che questo crea l'impossibilità di usare quella proprietà per addestrare il modello predittivo. Nella simulazione che stiamo facendo sulla storia dei dati, dovremo escludere questo nuovo insieme fino a quando, nel 2020, saranno disponibili anche questi dati.

Spero di non averti annoiato con questa storia dei dati! Era però importante segnalare quanti siano i problemi che dobbiamo affrontare quando usiamo i dati per addestrare un modello.

Sì, confesso che è stato piuttosto pesante seguirti, ma mi sembrano tutte precisazioni importanti…

4. Le minoranze hanno sempre torto

Il titolo un po' scherzoso vuole significare che se in un insieme di dati di addestramento che riguardano, ad esempio, detenuti, una etnia o categoria di persone è poco rappresentata, il modello predittivo tenderà a classificare queste persone un po' a caso….

E perché mai?

Guarda la Figura 9.7: in questo caso i dati rappresentano tanti detenuti Bianchi e Afro-Americani e un solo detenuto nativo americano. Nel metodo basato sul guadagno informativo, le caratteristiche da usare nell'albero per raggiungere un'alta precisione sono scelte sulla base di quanto riescono a discriminare i detenuti tra quelli che hanno commesso e non hanno commesso recidiva. Nel calcolo della entropia, e quindi del guadagno informativo, ogni detenuto conta uno, e perciò l'ottimizzazione sarà fatta sulle classi di dati numerose; le classi piccole saranno classificate un po' a caso, perché è irrilevante e casuale il loro contributo alla classificazione.

Chiaramente è necessario prendere contromisure per le classi poco rappresentate.

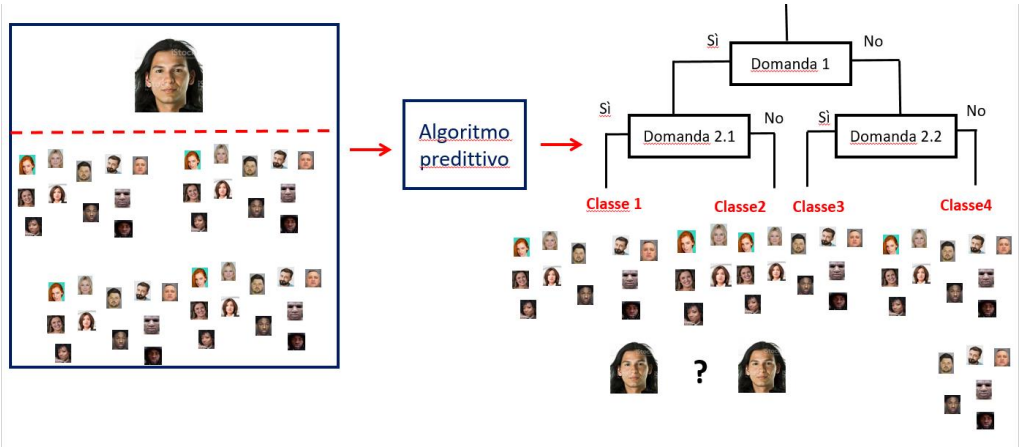


Figura 9.7 – Le minoranze hanno sempre torto …

5. Influenza della qualità dei dati

Fino ad ora abbiamo ragionato sull'esempio della concessione della libertà provvisoria come se vivessimo in un mondo perfetto, in cui tutti i dati sono accurati, completi, aggiornati.

Nell'esempio del paragrafo precedente sulla storia dei dati, abbiamo cominciato a vedere che le cose non stanno sempre così, e può accadere che i valori di qualche caratteristica non siano noti, per cui abbiamo scritto il valore *non so*.

Dedicherò un volume della Enciclopedia alla qualità dei dati, per ora ragioniamo su quanto vediamo nella Figura 9.8. Nella parte sinistra della figura compare la solita tabella di dati sui detenuti, in cui tutti i valori nella tabella sono da considerarsi *accurati* (cioè se per il detenuto 2 il numero di arresti precedenti è "uno o +", questo è *vero nella storia del detenuto*, e non è vero che non abbia avuto nessun arresto), *completi* (non c'è nessun non so) e *aggiornati* (per la caratteristica *reato contestato* viene considerato l'ultimo reato, anche nel caso in cui, come accade talvolta, il reato contestato sia stato modificato nel corso della reclusione).

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	D	grave	nessuno	No
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	uno o +	Si
8	D	grave	nessuno	No
9	D	grave	uno o +	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	leggero	nessuno	Si
13	A	grave	uno o +	SI
14	U	leggero	uno o +	Si

	Genere	Reato contestato	#Arresti precedenti	Recidiva?
1	U	leggero	Non so	Si
2	U	grave	nessuno	Si
3	U	Non so	Non so	No
~~4~~	~~D~~	~~grave~~	~~nessuno~~	~~No~~
5	D	leggero	uno o +	Si
6	U	leggero	nessuno	No
7	U	grave	nessuno	Si
~~8~~	~~D~~	~~grave~~	~~nessuno~~	~~No~~
9	D	grave	nessuno	Si
10	U	leggero	uno o +	Si
11	U	grave	nessuno	Si
12	U	grave	nessuno	Si
13	A	grave	uno o +	SI
14	U	grave	uno o +	Si

Figura 9.8 – Qualità dei dati: accuratezza, completezza, aggiornamento

Nella parte destra della figura vediamo alcun esempi di dati di scarsa qualità, dati che, cioè, non corrispondono al valore reale. Le ragioni della scarsa qualità possono essere molteplici.

Per esempio la cancellazione della riga 4 e della riga 8 può derivare da un problema di trasmissione o memorizzazione, i valori *nessuno* in rosso possono derivare da ritardati aggiornamenti (ad esempio il detenuto potrebbe aver trascorso periodi di pena in altri penitenziari, e non esiste una base di dati centralizzata che registra tutti gli arresti). I reati dei detenuti 12 e 14 sono erroneamente indicati come gravi, per un errore di digitazione o scelta in una finestra. Insomma, talvolta i dati sono sbagliati per ragioni banali, che però accadono con una certa frequenza.

Il problema è che se non ci si accorge che i dati sono sbagliati, la loro scarsa qualità può influire sul processo di apprendimento, perché i conteggi finali sui positivi e negativi sono influenzati da questi errori. E spesso accade che non ci si accorga dell'errore, perché per scoprire che un dato è sbagliato dobbiamo "inciampare sul dato", cioè ci accorgiamo dell'errore perché un programma che usa il dato fornisce un risultato manifestamente sbagliato.

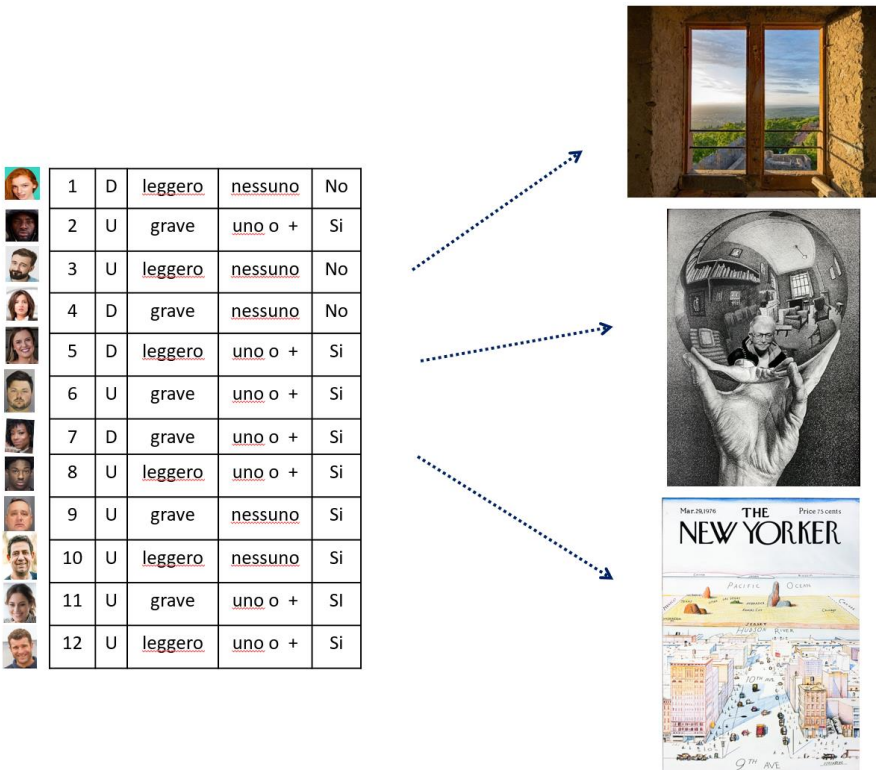
Potete fare per conto vostro un *esercizio*, per esempio partendo dalla tabella con 22 detenuti di Figura 9.5, potete cambiare alcuni valori con altri valori errati, e vedere come cambia il modello predittivo costituito dall'albero di decisione.

6. I dati possono essere uno specchio deformante della realtà

Il primo volume della Enciclopedia è partito da questa immagine: i dati sono una finestra sul mondo. Può accadere che questa finestra sia uno specchio deformato, e ci dia una immagine distorta del mondo. Questo è ciò che accade nella Figura 9.9 in cui vediamo nella parte centrale la famosa sfera di Escher con me riflesso, ovvero l'altrettanto famoso manifesto su New York, in cui man mano che ci si allontana dalla Quinta Strada l'immagine del mondo si fa via via meno precisa e dettagliata.

Questo accade spesso nella nostra vita: un evento drammatico, ad esempio un terremoto con tante vittime in Italia, ci colpisce molto di più dello stesso evento in un paese africano o asiatico. Questo accade anche nel nostro lavoro, ci sono particolari aspetti che ci attraggono più di altri. Può accadere perciò all'analista che raccoglie caratteristiche del fenomeno su cui costruire il modello predittivo; può anche accadere che *la lente deformata sia quella del modello predittivo*, per cui ci si può incaponire nel cercare caratteristiche che portano ad un elevato guadagno informativo, ma che non c'entrano niente con gli scopi del modello.

Propongo qui un esercizio che sviluppi questa idea con casi concreti: provate a indagare ad esempio caratteristiche o insiemi di caratteristiche rilevanti per il caso Compas che sono rappresentate con minor dettaglio e completezza di altre, meno rilevanti.



1	D	leggero	nessuno	No
2	U	grave	uno o +	Si
3	U	leggero	nessuno	No
4	D	grave	nessuno	No
5	D	leggero	uno o +	Si
6	U	grave	uno o +	Si
7	D	grave	uno o +	Si
8	U	leggero	uno o +	Si
9	U	grave	nessuno	Si
10	U	leggero	nessuno	Si
11	U	grave	uno o +	SI
12	U	leggero	uno o +	Si

Figura 9.9 – I dati sono una finestra sul mondo,
ma questa finestra può essere una finestra deformata

9.3 Il modello

Qui analizziamo i problemi etici insiti nella costruzione del modello predittivo.

1. I modelli e le tecniche non sono mai neutrali: Bias sociali e Bias statistici

I tre mondi che abbiamo già preso in considerazione e che riproduco in Figura 9.10, sono il contesto in cui i modelli predittivi vengono generati e utilizzati.

I bias, le distorsioni che portano a compromettere le equità del modello, si generano in entrambi i passaggi dal mondo ideale al mondo reale, e da questo al mondo dei dati digitali.

Per esempio, un bias verso le persone di una determinata etnia si manifesta nel primo passaggio, non c'entra nulla con quanto avviene nel mondo digitale. L'uso della traduzione nel genere maschile per doctor e del genere femminile per nurse dell'esempio nel Capitolo 1, non è dovuto al modello che effettua la traduzione, è dovuto al fatto che i testi a partire dai quali il modello impara a tradurre sono stati scritti da umani, sono scritti in lingue in cui il problema esiste da sempre.

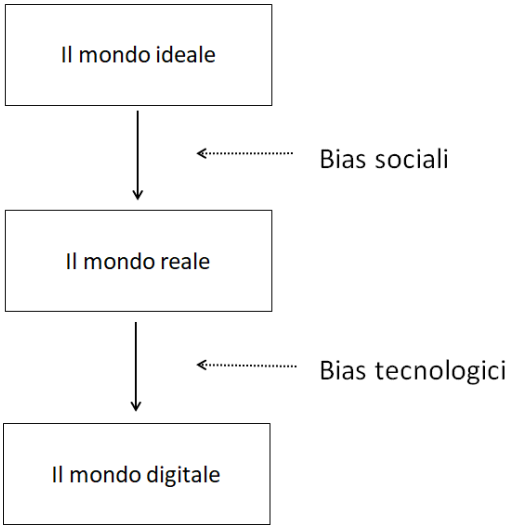


Figura 9.10 – I bias sociali e i bias tecnologici

Se invece abbiamo rappresentato una minoranza con pochi elementi e non ci siamo premurati di usare tecniche che permettano ugualmente di classificare questi elementi in modo accurato, questo è un bias di diversa natura, che dipende dal modo con cui abbiamo scelto le tecniche di apprendimento; questi bias sono diversi dai precedenti, e li chiameremo nel Capitolo 10 dedicato alla mitigazione *bias tecnologici*, per ragioni che chiarirò nel Capitolo.

2. Tradeoff tra accuratezza e spiegabilità: ovvero, non si può ottenere tutto dalla vita

In questo libro ci siamo occupati soprattutto della accuratezza ed equità dei modelli. Se però ritorniamo ai risultati della Conferenza Call for Ethics del Capitolo 5, vediamo che tra le qualità etiche citate nel documento finale troviamo la

Trasparenza: i modelli predittivi dell'Intelligenza Artificiale dovrebbero essere spiegabili, comprensibili

Nel seguito non userò il termine *trasparenza*, che può riguardare tanti ambiti delle tecnologie dei dati digitali, preferisco usare il termine *spiegabilità*. Abbiamo già iniziato a discutere della spiegabilità quando abbiamo visto che gli alberi di decisione sono decisamente più semplici da comprendere rispetto alle foreste casuali di alberi, mentre queste ultime sono superiori rispetto agli alberi per quanto riguarda l'accuratezza. Indaghiamo un po' più approfonditamente la relazione tra le due qualità.

Collochiamo gli alberi di decisione e le foreste causali in un diagramma, sulle cui coordinate riportiamo l' accuratezza e la spiegabilità, vedi Figura 9.11.

La spiegabilità ha lo scopo di rendere esplicite le interazioni tra la tecnica di apprendimento usata per produrre il modello, il modello e i dati su cui esso opera; essa è rilevante sia quando di un modello vogliamo comprendere il funzionamento, sia per scoprire discriminazioni sistematiche (gli Afro-Americani sono discriminati), sia quando si vuole spiegare l'esito del modello per un singolo individuo (perché Verdi non è stato liberato?).

Supponiamo, per esempio, che un modello produca una graduatoria per accedere a un servizio, per esempio una casa popolare. Se un individuo inserisce i suoi dati e riceve come risultato un punteggio, questo numero da solo non fornisce alcuna informazione sul perché sia stato assegnato tale punteggio e sul perché della posizione comparativa rispetto agli altri partecipanti. E' necessario capire *come* ha fatto la tecnica a generare quel punteggio.

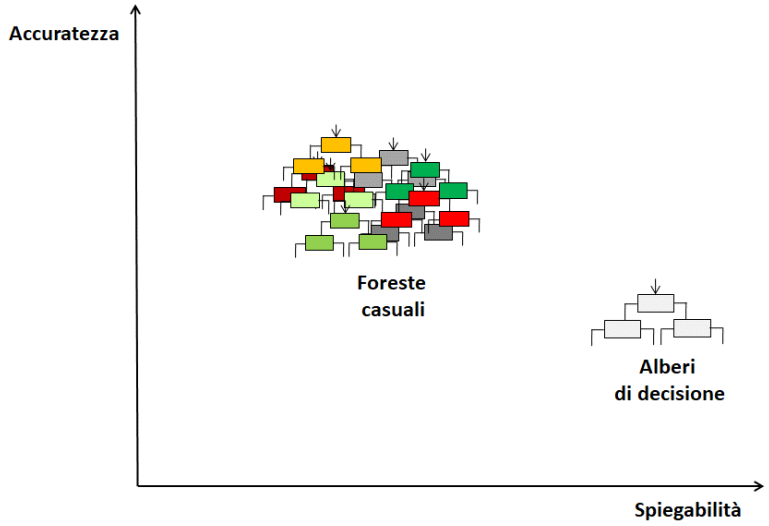


Figura 9.11 – Alberi di decisione e foreste causali nello spazio della accuratezza e spiegabilità

I precedenti esempi ci introducono alla distinzione tra *spiegabilità globale e locale*; una un modello è *globalmente spiegabile* se siamo in grado di comprendere la logica complessiva del modello e seguire l'intero ragionamento che porta ai differenti possibili risultati, ad esempio come il modello stabilisce globalmente per i detenuti il livello di rischio di recidiva; è *localmente interpretabile* se siamo in grado di comprendere le motivazioni per una specifica predizione, ad esempio perché un certo detenuto abbia rischio basso o alto, ovvero perché siamo arrivati decimi in una graduatoria di un concorso.

Tornando agli alberi di decisione, vediamo che l'albero esprime il modello predittivo attraverso un insieme di scelte basate su valori di caratteristiche: ad esempio, quando una scelta è sul genere, ci chiediamo se il detenuto sia uomo o donna, se la scelta è sul numero di arresti precedenti, ci chiediamo se siano 0 ovvero uno o più di uno. Chiarire bene il significato delle caratteristiche contribuisce a migliorare la spiegabilità. Un ultimo aspetto che fa capire "come ragiona un albero di decisione" sono la spiegazione dei guadagni informativi connessi alle diverse caratteristiche prese in considerazione.

Quando dagli alberi passiamo alle foreste, le cose diventano molto più complicate, come sappiamo se amiamo le passeggiate e amiamo (come me) perderci in un bosco. Certamente, conosciamo gli alberi "antenati" che sono utilizzati per il processo generativo, ma non riusciamo a comprendere la complessità della foresta, e soprattutto non riusciamo a comprendere gli aspetti legati alla casualità nella scelta delle caratteristiche e dei dati in ingresso. Certo, intuitivamente ci fidiamo di più di una foresta, perché intuiamo il fatto che le decisioni prese da molti alberi sono più affidabili delle decisioni prese da uno solo, ma è proprio questo fidarsi della foresta che è una prima fondamentale rinuncia a comprendere a fondo.

Potrebbe venirci una idea, guardiamo Figura 9.12. Se abbiamo prodotto un modello generativo che utilizza una foresta casuale, perché non proviamo a costruire un albero di decisione che produce un modello equivalente a quello della foresta, risolvendo così il problema della spiegabilità?

L'idea è buona, e infatti sono stati sviluppati modelli predittivi per trasformare una foresta in un albero di decisione, e tuttavia ha un limite di fondo che possiamo intuire guardando nuovamente la Figura 9.12: l'albero "equivalente" sarà sempre meno preciso della foresta, quindi è illusorio pensare di adottarlo *al posto della foresta*.

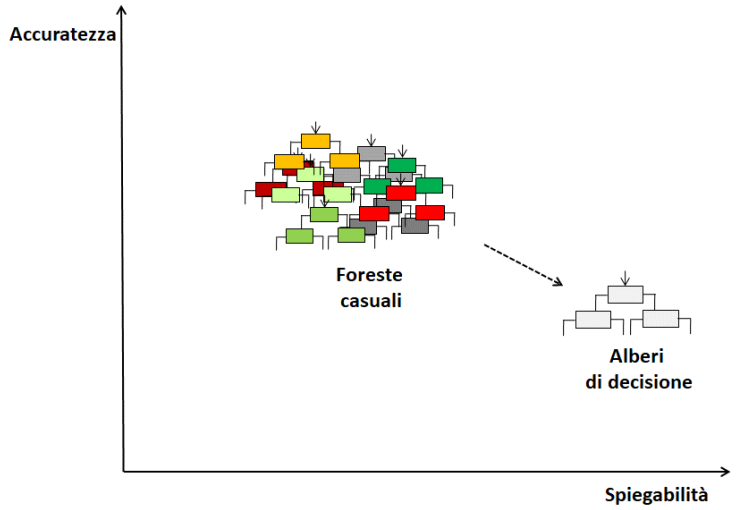


Figura 9.12 – Foresta causale e sua trasformazione in albero di decisione

Accanto agli alberi di decisione e alle foreste casuali di alberi, molte altre tecniche sono state proposte per costruire modelli predittivi. Tra esse vediamo un po' più da vicino le reti neurali,[14] modello matematico/informatico di calcolo basato sulle reti neurali biologiche. Il modello è costituito da un gruppo di interconnessioni di informazioni costituite da neuroni artificiali, e processi che utilizzano un approccio di calcolo di tipo *connessionistico*, vedi tra poco il significato.

Una rete neurale (vedi Figura 9.13) riceve segnali esterni su uno strato di unità di elaborazione o nodi d'ingresso, ciascuno dei quali è collegato con nodi interni, organizzati in più livelli. Ogni nodo elabora i segnali ricevuti e trasmette il risultato ai nodi successivi. La rete è una struttura che permette di simulare relazioni complesse tra ingressi e uscite che altre funzioni non riescono a rappresentare. I processi di calcolo sono basati sul connessionismo, il modello delle scienze cognitive, in cui il cervello umano elabora le informazioni ricevute dagli organi dei sensi in modo parallelo e distribuisce le informazioni in tutti i differenti nodi della rete cognitiva.

[14] Il seguito è tratto da Wikipedia, rete neurale, 2022.

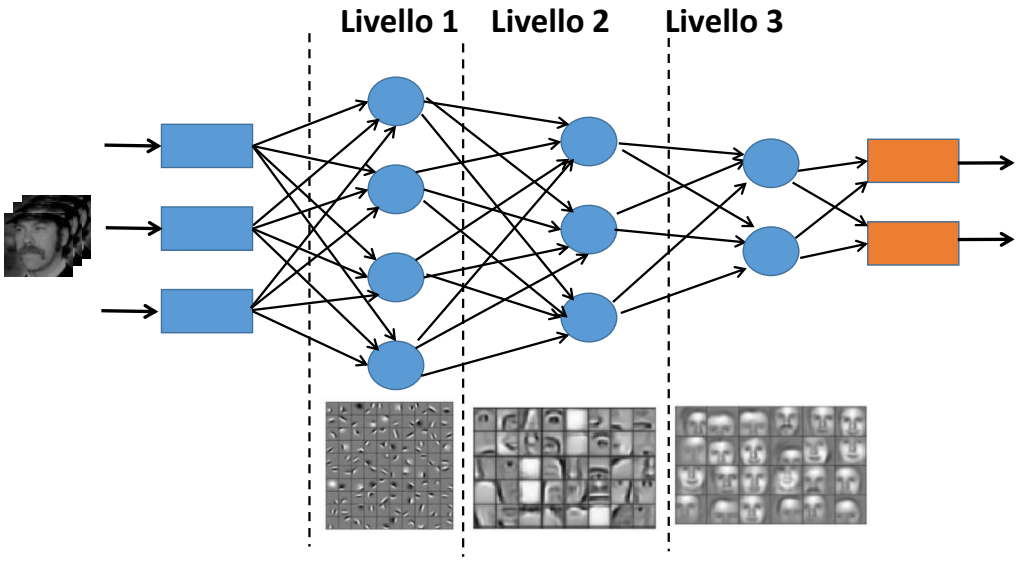


Figura 9.13 – Una rete neurale organizzata in tre livelli, che legge in ingresso immagini di volti di persone e riproduce in uscita un identificatore o una classificazione delle persone corrispondenti

Il processo di riconoscimento dei volti mostrato in Figura 9.13 procede secondo diverse fasi:

1. Estrazione di uno o più volti dalla immagine
2. Normalizzazione della immagine del volto rispetto a una immagine standard, ad esempio un volto visto di fronte con le due parti del volto simmetriche rispetto a un asse. Questa fase non è mostrata in figura.
3. Estrazione delle caratteristiche elementari del volto usate nelle successive fase del riconoscimento
4. Aggregazione delle caratteristiche elementari in parti del volto.
5. Riconoscimento del volto tra uno o più volti di una base di dati
6. Classificazione del volto come appartenente a una etnia o altra forma di classificazione.

Senza entrare in maggiore dettaglio sulla struttura delle reti neurali e su come si progettano, dico solo che i livelli intermedi sono chiamati spesso livelli nascosti (hidden layers) perché le operazioni che vengono svolte in questi livelli sono espresse da equazioni matematiche che trasformano i dati in input in quelli in output, non descrivibili in modo comprensibile. Mentre un albero di decisione separa l'insieme da classificare mediante predicati del tipo "è donna?", "è già stato incarcerata/o una o più volte?", che sono facilmente comprensibili, una rete neurale usa modelli di calcolo cosiddetti *nascosti*.

Questo è il limite delle reti neurali, la scarsa spiegabilità. Ma accanto a tale limite, le reti neurali possono efficientemente e con grande accuratezza risolvere problemi complessi su grandi quantità di dati per mezzo di un numero di livelli che cresce con la complessità del compiti e una grande varietà e potenza di operazioni di trasformazione.

Per tale ragione le reti neurali si collocano nel quadrante Accuratezza/Spiegabilità, accanto agli alberi di decisione e alle foreste causali, nella posizione mostrata in Figura 9.14.

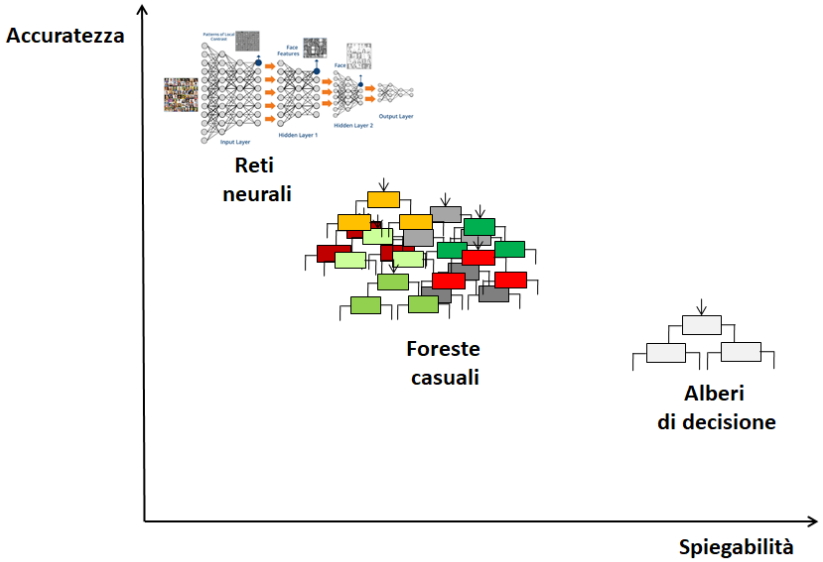


Figura 9.14 – Tradeoff tra accuratezza e spiegabilità

Ma spiegami bene, non c'è speranza di trovare tecniche che siano allo stesso tempo accurate e spiegabili?

Il tradeoff, cioè la relazione funzionale tra accuratezza e spiegabilità, per cui quando l'una aumenta l'altra diminuisce è diffuso nella scienza; forse l'esempio più noto è il principio di indeterminazione di Heisenberg, per cui non è possibile misurare contemporaneamente e con esattezza le proprietà che definiscono lo stato di una particella elementare, la sua velocità e la sua posizione. In questo caso la ricerca procede scoprendo sempre nuove tecniche che, come mostrato graficamente in Figura 9.15, migliorano le tecniche nelle stesse famiglie.

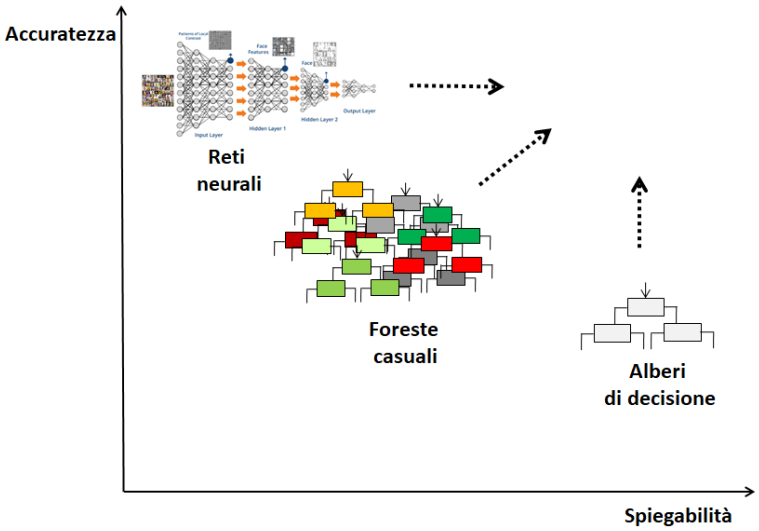


Figura 9.15 – Le tecniche scoperte in nuove ricerche migliorano la coppia di qualità accuratezza/spiegabilità

Comunque se ne vuoi sapere di più ti consiglio di leggere il lavoro riportato in nota[15].

9.4 Il feedback

Accade spesso che i modelli predittivi facciano uso delle reazioni degli utenti come retroazione (feedback) da utilizzare per affinare il modello di predizione. Vediamo alcune situazioni in cui è importante rendersi conto della influenza dei feedback.

1. Quanto possono essere ambigui i click!

Un modello predittivo applicato nel mondo reale, lo modifica. Concedere o non concedere la libertà provvisoria a un detenuto determina un futuro diverso nella vita del detenuto e dei suoi familiari.

Per capire l'influenza di tali decisioni sulla vita degli individui è necessario svolgere indagini sociali complesse. In altre circostanze, però, l'esito del modello predittivo può diventare immediatamente un nuovo input al modello stesso. Ad esempio, un sistema di raccomandazione che cerca di prevedere e orientare le esigenze di acquisto di un cliente sottoponendogli indirizzi di siti, può avere un immediato riscontro del successo delle proprie raccomandazioni dai click che il cliente ha effettuato per selezionare le pagine dei siti. Oppure, un modello predittivo per il gioco del Go che individua la mossa che dà maggior vantaggio, può modificare la propria strategia sulla base della mossa successiva dell'avversario. Si parla in questo caso di *apprendimento per rinforzo*, o reinforcement learning, che abbiamo visto nel Capitolo 4.

Alcune volte però il click dell'utente può essere ambiguo. Tornando al sistema di raccomandazione, il fatto che l'utente selezioni la prima pagina proposta dal sistema non significa necessariamente che la raccomandazione sia stata accolta, può anche accadere che l'utente sia abituato a selezionare sempre la prima pagina per abitudine o per pigrizia.

Insomma, dobbiamo essere consapevoli che le nostre reazioni sono prese in considerazione molto seriamente dai modelli predittivi, in particolare dai sistemi di raccomandazione: ma possono essere anche interpretate con significato diverso rispetto alla nostra intenzione.

Mi è capitato di fantasticare su come ingannare un sistema di raccomandazione, selezionando, ad esempio prodotti che non comprerei mai. Ma bisogna stare attenti: un mio collega che doveva prendere un aereo per Bari si sbagliò nel digitare la lettera r e scrisse *Bali*; da qual momento fu tempestato per molto tempo da banner pubblicitari di agenzie che vendevano viaggi a Bali e da pubblicità su Bali. Non ci fu verso di far cambiare idea al sistema di raccomandazione…

2. Le predizioni influiscono sul comportamento degli utenti

[15] R. Guidotti A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi - A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys, 2018.

Siete a Roma e dovete andare da piazza Cavour, sede del Ministero di Giustizia, alla Stazione Termini. Siete abituati a fare questo itinerario passando dal centro, Via del Tritone, Via Volturno e infine Termini.

Questa volta guardate il navigatore, vedi Figura 9.16, che vi propone di passare dal Muro Torto, Castro Pretorio e Via Marsala. C'è un po' di traffico sul Lungotevere e a via Marsala, ma ci mettete di meno. La reazione quasi sicura e che vi fidate, e passate dal Muro Torto.

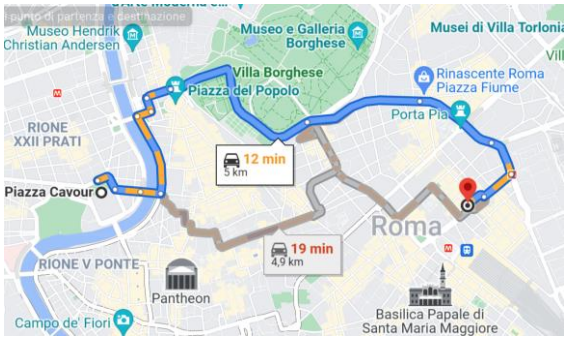


Figura 9.16 - le previsioni influiscono sul comportamento degli utenti

I modelli come il navigatore, che vi predicono quanto tempo ci metterete a fare un certo itinerario, influenzano il comportamento degli utenti, ma se tutti gli utenti accettano il suggerimento, l'itinerario in centro tenderà a diventare più fluido e veloce, e quello dal Muro Torto tenderà ad assorbire più traffico e a diventare più lento.

Questo esempio ci dice che i modelli predittivi come visto nella Figura 9.16 operano in un mondo popolato da esseri umani, e questi esseri umani possono influire consciamente o inconsciamente sui fenomeni che a loro volta influenzano e modificano i modelli.

Questa, secondo te, è una buona o una cattiva notizia?

Come spesso accade è sia buona che cattiva. E' cattiva se noi siamo fruitori inconsapevoli dei risultati dei modelli predittivi, è buona se, consapevoli di essere quello che viene chiamato lo *human in the loop*, l'essere umano nel ciclo, ci rendiamo conto che possiamo esercitare un ruolo attivo nella interazione con il modello, informandoci sulla sua qualità e magari anche sul processo di apprendimento che ne genera le predizioni.

3. La profezia che si autoavvera

Questo detto: la profezia che si autoavvera, viene usato per indicare situazioni in cui viene formulata una previsione (ad esempio, sento che domani avrò una discussione difficile sul lavoro), e gli eventi e i comportamenti delle persone vanno tutti nel senso di farla accadere (entro in uno stato di ansia che mi porta ad essere nervoso e a creare un clima teso).

La profezia che si autoavvera si verifica anche nei modelli predittivi. Consideriamo la Figura

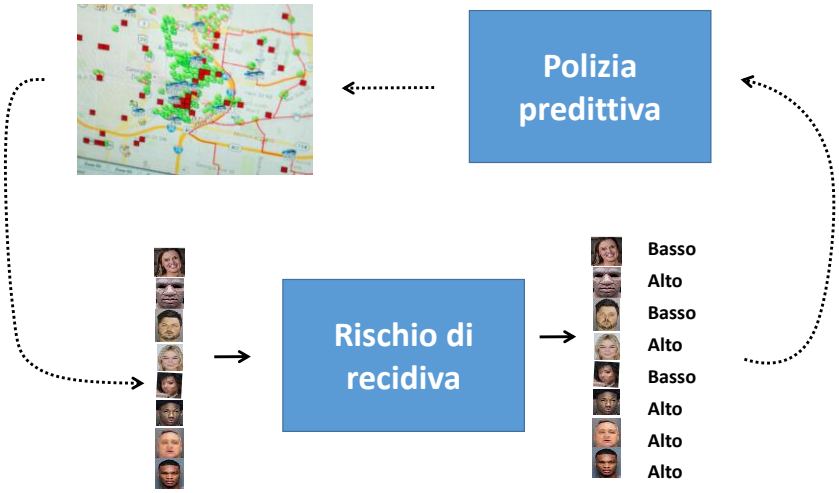


Figura 9.17 – La profezia che si autoavvera

9.17 in cui l'usuale modello predittivo sul rischio di recidiva viene usato insieme ad un modello di polizia predittiva o predictive policing, molto usato negli Stati Uniti, in cui il modello prevede in quali zone di una città e in quali orari è più probabile che si verifichino certi reati.

L'uso congiunto dei due modelli può spingere la polizia di una città a fare questo ragionamento: siccome il modello sul rischio di recidiva dice che i soggetti più a rischio sono gli Afro-Americani, io pattuglio di più i quartieri con maggioranza di cittadini Afro-Americani. In questo modo *vengono rilevati più reati tra gli Afro-Americani*, che rafforzano il modello previsionale di predictive policing e aumentano in percentuale gli Afro-Americani che vengono fermati e spediti in prigione, con un rafforzamento reciproco tra i due modelli, e una previsione di rischio per gli Afro-Americani che viene via via confermata.

9.5 Quando è pericoloso usare i modelli: le indicazioni della Unione Europea

Nel 2021 l'Unione Europea ha definito un piano di azione coordinato in tema di Intelligenza Artificiale, delineando un insieme di linee guida[16] per i paesi componenti e per le aziende del mercato. Vediamo le principali tra quelle attinenti questo libro, che si concentra su un aspetto importante ma specifico nella Intelligenza artificiale, quello dei modelli basati su apprendimento automatico.

La prima linea guida è veramente "controcorrente"; a fronte di molte posizioni che magnificano i vantaggi della Intelligenza Artificiale (IA) e, nel nostro contesto, del Machine learning, l'Unione Europea propone di *non usare* i modelli classificatori e predittivi in alcune aree considerate molto pericolose. L'elenco delle pratiche vietate comprende tutti i sistemi di IA il cui uso è considerato inaccettabile, in quanto contrario ai valori dell'Unione, ad esempio quando violano i diritti fondamentali; esse sono mostrate in Figura 9.18.

[16] Le linee guida non sono ancora una direttiva che i Paesi devono obbligatoriamente rispettare.

Non si possono usare, ad esempio, sistemi di raccomandazione quando usano tecniche che sollecitano la mente umana sotto il livello della coscienza, tecniche che sono troppo *deboli* per essere avvertite razionalmente, e che però riescono a influenzare l'inconscio e condizionare il comportamento. Non si può rischiare che i modelli provochino danni mentali o fisici. I modelli predittivi non possono essere usati per stabilire graduatorie su servizi sociali.

- Pratiche che hanno un potenziale significativo di manipolazione delle persone attraverso tecniche subliminali al di là della loro coscienza.
- Sfruttamento delle vulnerabilità di soggetti come bambini o persone con disabilità, al fine di distorcere materialmente il loro comportamento in modo tale da causare loro o un'altra persona danni psicologici o fisici.
- Graduatorie sociali per scopi generali da parte delle autorità pubbliche.
- Uso di sistemi di identificazione biometrica "in tempo reale" in spazi accessibili al pubblico a fini di contrasto, a meno che non si applichino determinate
- eccezioni.

Figura 9.18 – In quali aree la Unione Europea consiglia di non usare i modelli basati su ML

Il regolamento segue un approccio basato sul *rischio*, differenziando tra usi della IA che creino a. un rischio inaccettabile, b. un rischio elevato e c. un rischio basso o minimo. Le aree che vengono viste come ad alto rischio sono[17]:

1. Identificazione biometrica e categorizzazione delle persone fisiche
2. Gestione e funzionamento delle infrastrutture critiche - sistemi destinati ad essere utilizzati come componenti di sicurezza nella gestione e nel funzionamento della circolazione stradale e nella fornitura di acqua, gas, riscaldamento ed elettricità.
3. Istruzione e formazione professionale
4. Occupazione, gestione dei lavoratori e accesso al lavoro autonomo.
5. Accesso a o fruizione di servizi pubblici o privati
6. Forze dell'ordine
7. Gestione delle migrazioni delle richieste di asilo e del controllo delle frontiere.

Il regolamento per ciascuna di queste aree entra nel dettaglio identificando le sottoaree, ad esempio per l'area 7:
a) sistemi utilizzati dalle autorità pubbliche competenti per rilevare lo stato emotivo di una persona;
b) sistemi utilizzati per valutare un rischio, compreso un rischio per la sicurezza, un rischio di immigrazione irregolare o un rischio per la salute, rappresentato da una persona fisica che intende entrare o è entrata nel territorio di uno Stato membro;
c) i sistemi utilizzati per la verifica dell'autenticità dei documenti di viaggio e della documentazione giustificativa delle persone fisiche e per rilevare i documenti non autentici verificandone le caratteristiche di sicurezza;

[17] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions su "Fostering a European approach to Artificial intelligence Proposal of a regulatory framework in AI and revised coordinated plan on AI", 21 aprile 2021.

d) sistemi utilizzati nell'esame delle domande di asilo, visti e permessi di soggiorno e relativi reclami per quanto riguarda l'ammissibilità delle persone fisiche che richiedono uno status.
e) Sistemi destinati ad assistere una autorità giuridica nella ricerca e interpretazione di fatti e delle leggi e nell'applicare la legge a insiemi di fatti riferiti al procedimento penale o al processo civile.

In tutti questi casi viene proposto di procedere nella progettazione di modelli predittivi attraverso un metodo che va nella direzione della riduzione del rischio, definendo alcune buone pratiche di progettazione, che riguardano i seguenti aspetti:

1. Uso di dati di addestramento che siano di qualità, e quindi accurati, completi e aggiornati.
2. Documentazione della tracciabilità nel funzionamento del modello, così da essere in grado di ricostruire le diverse fasi del processo di apprendimento. Questo obiettivo si avvicina molto alla caratteristica di qualità che abbiamo chiamato spiegabilità.
3. Coinvolgimento dell'utente attraverso condivisione di conoscenza e sensibilizzazione alla consapevolezza dei vantaggi e dei danni che possono essere causati da un utilizzo non equo dei modelli.
4. Raggiungimento di obiettivi di robustezza, accuratezza, sicurezza (proprietà RAS). La *robustezza* si acquisisce quando il modello è resistente rispetto a possibili guasti o uso con dati anomali e di scarsa qualità. Sulla accuratezza ho parlato diffusamente nel Capitolo 7. La *sicurezza* risponde all'obiettivo di contrasto rispetto ad attacchi al funzionamento del modello da parte di soggetti esterni.
5. Collaudo di conformità rispetto alle proprietà RAS. Questo significa che prima di essere utilizzato, il modello deve essere sottoposto a test di funzionamento che assicurano le proprietà RAS.
6. Approccio alla progettazione che adotti regole tecniche le quali non creino eccessivo carico burocratico inutile verso gli operatori economici. Questo è un punto molto importante: quando si emette una norma, occorre sempre considerare quanto la norma raggiunga il suo obiettivo, senza la necessità di provocare costi per le aziende o sacrifici inutili agli utenti.

Come già detto in nota, le norme europee sono per il momento (2022) una raccomandazione e sono destinate a diventare una direttiva, ponendo la Unione Europea nella scia di quanto accaduto per il GDPR, il Regolamento per i dati personali, e cioè all'avanguardia nei processi di regolazione sulle tecnologie informatiche e della Intelligenza Artificiale.

9.6 Il cosa e il perché

Ho brevemente introdotto nel Capitolo 2 i modelli interpretativi, spiegati con l'esempio del mal di testa. Credo capiti a tutti certi giorni di essere nervosi, insofferenti. In quei giorni ci chiediamo: perché sono nervoso? Mi è capitato qualcosa, ho avuto una brutta notizia? E' la giornata nuvolosa e piovosa? E' quel cielo latteo? Non riesco a fare una certa cosa? Sono insoddisfatto? Sono tutte queste cose insieme?

Oppure, certe volte ci capita di pensare sulle ingiustizie nel mondo, e anche in questo caso non ci fermiamo solo a osservarle, ma cerchiamo di capire quali sono le cause. I bambini

piccoli chiedono spesso alla mamma e al papà: perché sono inciampato? Perché l'insetto mi ha pizzicato? Perché sento dolore alla mano?

Spesso nella vita cerchiamo le cause delle cose, degli avvenimenti, degli stati d'animo. C'è un libro bellissimo che ragiona su questo grande tema, l'autore è Judea Pearl e il libro si chiama il *Libro del perché*, The Book of Why. Anche io potrei dedicare un intero libro di questa Enciclopedia al tema del *perché*, ma per ora ne vedremo brevemente alcuni importanti aspetti trattati da Pearl, legati al tema etico dei modelli predittivi basati su Machine learning.

Nel 2008 Chris Anderson, direttore della rivista Wired, scrisse un editoriale dal titolo 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', ovvero "La fine della teoria: il diluvio dei dati rende il metodo scientifico obsoleto", creando un ampio dibattito sulla grande novità portata nel metodo scientifico dalla disponibilità di quantità di dati massive, i big data.

Se ci pensiamo bene, il Machine learning, come lo abbiamo investigato il questo libro, non si pone mai la domanda sulle *cause* dei fenomeni. Per prevedere quale è il grado di rischio che un detenuto liberato commetta recidiva, non ci poniamo mai la domanda del perché sia detenuto in un istituto di pena, ma lo confrontiamo con tutti i detenuti liberati nel passato, assegnandogli un livello di rischio proporzionato a quello dei detenuti liberati nel passato che sono più simili a lei/lui.

Questo lavoro, consistente nello scoprire il livello di rischio e nell' individuare le similitudini tra detenuti del passato e detenuti nuovi, lo affidiamo a un modello predittivo che viene costruito senza porsi mai il problema delle cause dei comportamenti nella storia delle persone.

La situazione è simile a quella di Figura 9.19. in cui un gufo osserva il comportamento dei topolini che si muovono ogni giorno da un posto A a un posto B. Il gufo non capisce la ragione di questo comportamento, ma osserva una *regolarità*, come una persona che fin da bambino osserva le eclissi di sole, e ricorda quando le eclissi si sono verificate, e trova delle regolarità in questo fenomeno negli anni.

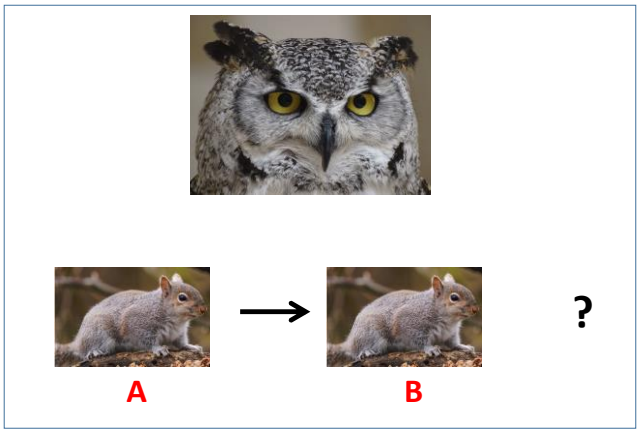


Figura 9.19 – Il Gufo non sa perché il topo va da A a B..

Anche Pearl parte da un esempio molto semplice. Pensiamo a un robot che è stato istruito a usare una aspirapolvere sulla base dei comportamenti di persone che hanno usato la aspirapolvere nel passato, vedi Figura 9.20.

Sono stati registrati dati dettagliatissimi sull'ordine con cui vengono aspirate le stanze, sulle parti delle stanze via via aspirate, sulla pendenza del bastone della aspirapolvere sotto i tavoli e mobili, ma nella rilevazione ci si è scordati di registrare l'ora del giorno in cui l'aspirapolvere viene usata, e l'intervallo temporale in cui il regolamento del condominio vieta di usarla.

L'aspirapolvere robot nel primo giorno di lavoro decide, per farsi apprezzare, di iniziare a utilizzare l'aspirapolvere molto presto, per lui il tempo non esiste, il tempo non è coinvolto nel *cosa* fare; ma non ha fatto i conti con la persona in fondo al disegno, che si sveglia presto per il rumore e protesta violentemente: perché mi hai svegliato!

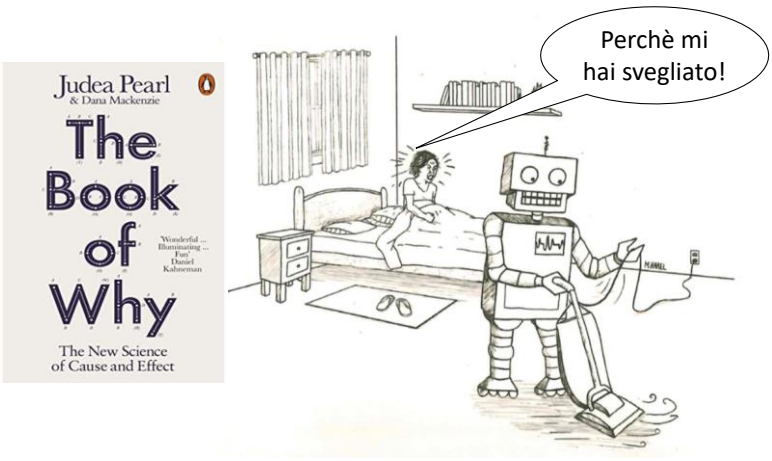


Figura 9.20 – I modelli predittivi che imparano dal passato…

Quella persona ragiona in termini del *perché*, non si capacita del comportamento del robot, *non ne capisce le ragioni*. E' molto chiaro, in questo esempio, il grande contrasto tra il cosa e il perché!

Per capire il perché non basta osservare il passato, dobbiamo *intervenire sulla realtà*; soltanto *intervenendo sulla realtà possiamo sperare di comprenderla*. Per esempio, per capire perché il topo vada ogni giorno da A a B, possiamo fare un esperimento, vedi Figura 9.21.

Mentre il topo sta nel posto A, mettiamo del parmigiano nel posto B, e osserviamo il suo comportamento. E' molto probabile che il topo corra verso B, salvo il fatto che stia dormendo e quel giorno non sia molto attivo (anche questo è un perché, in questo caso motiva perché non si muove). Se ripetiamo molte volte l'esperimento, e vediamo che il topo va sempre da A a B nel momento in cui mettiamo in B del parmigiano, e non ci va quando mettiamo della carne, possiamo indurre un comportamento universale: il topo va da A a B *perché* è goloso di parmigiano. Questo ragionamento è chiamato *causale*.

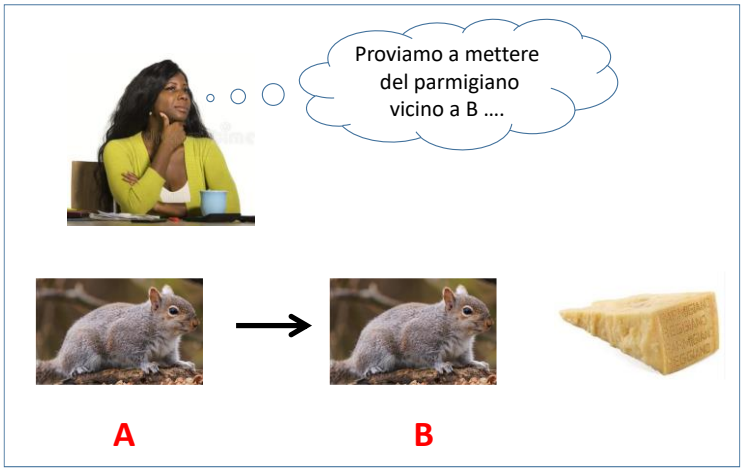


Figura 9.21 – Intervenire sulla realtà: proviamo a fare un esperimento con il formaggio…

Tutto chiaro. Però, ho una domanda: se non riescono mai a comprendere il perchè delle cose, come mai i modelli predittivi hanno tanto successo, sono tanto usati?

La ragione è abbastanza semplice, concentriamoci su un modello predittivo usato da un sistema di raccomandazione, per esempio quello di Amazon, che ti consiglia di comprare libri acquistati da altre persone simili a te: dal punto di vista del direttore vendite bastano le osservazioni e le associazioni create dalle tecniche predittive, non è necessario indagare perché le persone comprano libri, in questo caso ha ragione Anderson, basta individuare le associazioni, *buone predizioni non devono necessariamente avere buone spiegazioni*.

C'è un terzo livello, dopo l'osservazione della realtà e l'intervento sulla realtà, che dobbiamo raggiungere per affrontare il problema del perché, quello dei *controfattuali*. Mi spiego con l'esempio dei vaccini.

Se voglio indagare l'effetto di un vaccino su una popolazione devo rispondere alla domanda: cosa accadrebbe se non prescrivo quel vaccino? Quindi, nel mio esperimento sul mondo, devo definire due gruppi di pazienti, uno a cui prescrivo il vaccino e l'altro a cui non lo prescrivo, il cosiddetto *gruppo di controllo*. Solo confrontando i due mondi io sono in grado di capire se un vaccino provoca, *causa* l'immunità dal virus per cui è stato concepito. Questo ragionamento è chiamato *controfattuale*.

E' vero, è proprio così! Senti. ho una domanda che ritengo importante e che mi risulta difficile formulare. Ci provo: se penso all'etica, se penso a come si può valutare eticamente se sono buoni o cattivi il comportamento e le azioni delle persone, delle comunità, delle nazioni, mi sembra importante indagare i motivi che li spingono ad agire in un certo modo. E se analizzo il mio comportamento, per riuscire a capire retrospettivamente se in una certa azione mi sono comportato bene o male, devo essere capace di osservare me stesso; i modelli basati su tecniche di apprendimento, sono in grado di osservare se stessi? Sono in grado di distinguere il bene dal male?

Fai una domanda di straordinaria rilevanza, che fa tremare le vene e i polsi! E mi scuso se ad essa dedico poche frasi, per un problema che è investigato dalla filosofia etica e da tutte le

religioni. Mi affido alle parole di Judea Pearl. Pearl si pone quattro domande, e fornisce risposte, che, avverte, sono ancora provvisorie e tracciano soprattutto un percorso di ricerca, di cui si vede ancora vagamente il traguardo finale….

1. Abbiamo già realizzato macchine che pensano?

Non abbiamo ancora realizzato "macchine" che pensano, in qualunque modo si interpreti questo concetto. Finora il pensiero umano è stato imitato in domini molto rilevanti ma specifici, come gli scacchi (il modello di Alpha Zero) o il Gioco del Go (il modello Alpha Go), o la sintesi di proteine (il modello Deep Mind). E ciò non è una sorpresa, perché è stato ottenuto per la capacità delle macchine di fare tanti calcoli velocemente, come richiesto dall'uso di big data nel Machine Learning.

2. Potremo realizzare in futuro macchine che pensano?

Molto probabilmente si, anche se è difficile prevedere in quali direzioni si orienterà la produzione di tecnologie basate sulla Intelligenza Artificiale. Storicamente, gli scienziati raramente *non hanno realizzato tecnologie* che erano in grado di produrre sulla base di risultati scientifici.
I grandi passaggi sono due:

- il concetto di *agenzia.*
- La coscienza di sé.

Una *agenzia* è un insieme di agenti autonomi che sono in grado di interagire tra loro utilizzando strategie basate sul ragionamento causale e sul ragionamento controfattuale. In questo modo, diversi agenti possono interagire, simulando il meccanismo di funzionamento della mente umana. Quanto alla coscienza di sé, il presupposto per una "macchina morale" è la capacità di riflettere sulle proprie azioni, che è insita nel ragionamento controfattuale.

Per sviluppare in una macchina la autoconsapevolezza, è necessario che la macchina riesca a rappresentare *un modello di sè stessa*, e una memoria delle proprie azioni passate. Realizzata la capacità di autoanalisi, seguiranno altre capacità come la empatia, il senso di equità, il libero arbitrio, l'autocontrollo, la capacità di elaborare strategie e previsioni a lungo termine, che ricadono anche esse nel ragionamento controfattuale.

3. E' giusto realizzare macchine che pensano?

Si intrecciano qui questioni etiche sulla opportunità di sviluppare tecnologie senza avere la certezza che esse *non verranno utilizzate per fini non etici*. Un punto di svolta è stata la bomba atomica su Hiroshima, da allora molti pensano che quella tecnologia non doveva essere prodotta.

4. Possiamo realizzare macchine che sono capaci di distinguere il bene dal male?

La risposta è: probabilmente sì, se doteremo le macchine delle stesse abilità cognitive che noi abbiamo, come l'empatia, l'autocontrollo, l'intuizione, così da permettere alle macchine di

prendere autonomamente decisioni, e non solo per imitazione, come accade nel Machine learning.

E probabilmente saremo capaci di apprendere come le macchine elaborino il libero arbitrio, e come le macchine saranno capaci di nascondere a noi umani i loro segreti. A questo punto potrebbe accadere che le macchine saranno in grado meglio di noi di distinguere il bene dal male, siano migliori di noi. E questo, conclude Pearl, sarà il più bel dono che la IA potrà fare alla Umanità.

Un interessante esercizio è quello di ragionare sulle affermazioni di Pearl, cercando dalla lettura di articoli e libri trovati nel Web, gli elementi per formarci un nostra opinione.

Conclusioni

Quale etica rispettano i modelli basati sul Machine learning? Quali aspetti dobbiamo indagare per determinare un comportamento etico nei modelli basati sul Machine learning? Queste domande sono tanto più importanti perché spesso i modelli basati sul Machine learning riguardano le persone, e determinano classificazioni e previsioni molto rilevanti per la vita delle persone.

Alcune questioni riguardano i dati di addestramento sottoposti alla tecnica di apprendimento, altre riguardano la stessa tecnica, che vogliamo accurata e spiegabile, per poterci fidare, e per poter capire come funziona. Altri infine riguardano il ruolo dei feedback sui comportamenti degli esseri umani e dei modelli.

Per fortuna viviamo in Europa, le cui istituzioni sono all'avanguardia nelle leggi e regolamenti per la tutela della privacy e per un uso consapevole e responsabile della Intelligenza Artificiale e dei modelli predittivi. Istituzioni che ci dicono: certe volte è meglio non usarli i modelli, troppo pericoloso per gli esseri umani. E certe altre volte ci dicono: quando l'uso dei modelli riguarda aree critiche del rapporto tra stato e privati con i cittadini, usa delle precauzioni nel progettare il modello e nell'usarlo, sii consapevole di ciò che fai.

I modelli classificatori e predittivi non bastano, certe volte, a soddisfare la nostra curiosità e il nostro voler conoscere le cause dei fenomeni: il perché. Per fortuna che Judea Pearl e altri ricercatori sono andati più avanti rispetto ai modelli basati su apprendimento, modelli che, in quanto basati sul Machine learning, non si pongono il problema di cosa ci sia dietro alle classificazioni e alle predizioni.

Solo con il ragionamento causale e con il ragionamento controfattuale le macchine cosiddette intelligenti potranno un giorno, forse, distinguere il bene dal male. Sarà una bella sfida con gli umani, che certe volte brancolano un po' nel buio.

Capitolo 10. Metodi per mitigare l'iniquità

10.1 Mitigare le iniquità: i ruoli della società (cioè tutti noi) e dei modelli e tecniche di Machine Learning

Supponiamo di aver prodotto un modello classificatorio per decidere la graduatoria di un concorso pubblico sulla base dei risultati ottenuti dai concorrenti nei concorsi precedenti; abbiamo verificato dai dati di test che la graduatoria non rispetta una delle definizioni di equità più utilizzate e più citate, la *parità demografica*, insomma hanno vinto in proporzione alla popolazione più uomini che donne. Ho due domande:

Domanda 1 - Siamo d'accordo che sia la parità demografica la forma di equità che vogliamo rispettare?

Domanda 2 - In caso positivo, come dobbiamo modificare il modello per fare in modo che sia rispettata la parità demografica?

Secondo esempio. Nel caso della libertà provvisoria e del rischio di recidiva, supponiamo che sia stata istituita una commissione di inchiesta sul fenomeno; le conclusioni della commissione sono state simili a quelle di ProPublica: i modelli predittivi sfavoriscono i detenuti di una certa etnia rispetto alla equità che abbiamo chiamato *percentuale di falsi positivi*. Insomma, quei detenuti hanno maggiore probabilità di essere considerati a rischio alto, e di non commettere successivamente recidiva; come possiamo procedere per ristabilire questa forma di equità? Questo è il tema del presente Capitolo 10, Metodi per mitigare la iniquità.

Riconsideriamo in Figura 10.1 il "modello del mondo" che abbiamo introdotto in precedenza; abbiamo distinto:

- il mondo come dovrebbe essere, il mondo ideale,
- il mondo come è, il mondo reale,
- il mondo come è rappresentato nei dati digitali, il mondo digitale;

sappiamo già che i bias (distorsioni) manifestati dai modelli sono in parte un retaggio dei *bias sociali*, impliciti nella Domanda 1 qui sopra. Quanto ai bias dei modelli, occorre ora introdurre un nuovo concetto, che riassume al suo interno tutto ciò che abbiamo discusso in questo libro in tema di modelli, tecniche e dati digitali, il concetto di *tecnologia*.

Con il termine *tecnologia* intendiamo[18] l'insieme di tecniche, competenze, modelli, metodi, processi e dati digitali usati nella produzione di beni e servizi o nel raggiungimento di obiettivi, come ad esempio la indagine scientifica.

[18] Definizione ripresa e adattata da Wikipedia inglese

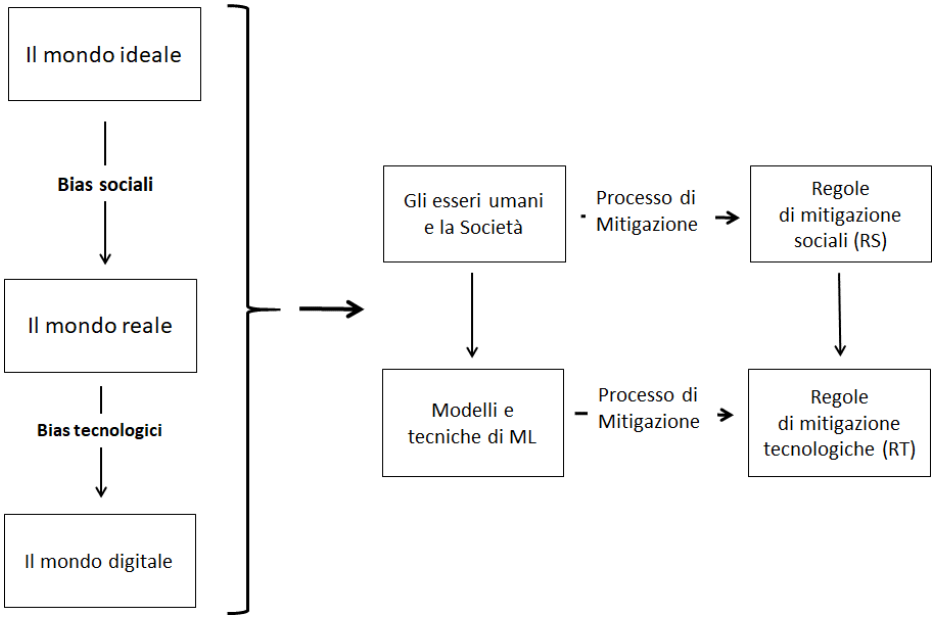


Figura 10.1 – Discriminazioni nella società e nel mondo digitale, mitigazioni mediante regole sociali e regole tecnologiche

La tecnologia può consistere nella *conoscenza* di tecniche, processi e simili, ovvero può essere *integrata in macchine*, al fine di permettere il loro utilizzo senza una conoscenza dettagliata del loro funzionamento. I sistemi che applicano la tecnologia trasformando un input in accordo alla funzione svolta dal sistema, e producendo un risultato, sono chiamati *sistemi tecnologici*.

I modelli classificatori e predittivi, le tecniche di Machine learning per la loro generazione, i dati digitali utilizzati dai modelli sono un particolare e sempre più diffuso tipo di *tecnologia*. Per questa ragione uso in Figura 10.1 il termine *bias tecnologici*; la Domanda 2 fa riferimento a questi.

Senti, perché hai introdotto ora questo concetto di tecnologia, che mi pare un po' fuori tema?

Scherzi, fuori tema?! L'Intelligenza artificiale viene applicata in tantissimi ambiti della nostra vita, dalla medicina alle armi da guerra, è fondamentale considerarla come una tecnologia, anzi, come la principale tecnologia che verrà utilizzata nel futuro!

Il nostro compito in questo capitolo è di indagare anzitutto su misure o *regole di mitigazione* che operano sui *bias tecnologici*, vedi Figura 10.1. Non dobbiamo considerare queste misure di mitigazione sufficienti: ce ne sono altre che dipendono da noi, che dipendono dalla sensibilità e dalle scelte nella società e nella politica in merito ai *bias sociali*, vedi ancora Figura 10.1. Indagheremo prima le regole sulle tecnologie e successivamente le regole di mitigazione sociali.

Gli esempi che vedremo in questo capito riguardano soprattutto modelli classificatori, in cui intendiamo suddividere la popolazione in ingresso in due insiemi, tipo: supera l'esame/non supera l'esame, oppure: è ammesso alla Università/non è ammesso alla Università. Le considerazioni generali e le regole si applicano ai modelli sia classificatori che predittivi.

10.2 – Regole di mitigazione tecnologiche (Regole T)

La situazione di partenza è questa: abbiamo prodotto un modello classificatorio o predittivo, abbiamo fatto misure di accuratezza sui dati di test, e abbiamo anche deciso di misurare *una o più delle tipologie di equità* introdotte nel Capitolo 8. Abbiamo ottenuto dei valori che dimostrano una forma di discriminazione verso un gruppo di elementi del campione dei dati di test, vedi Figura 10.2.

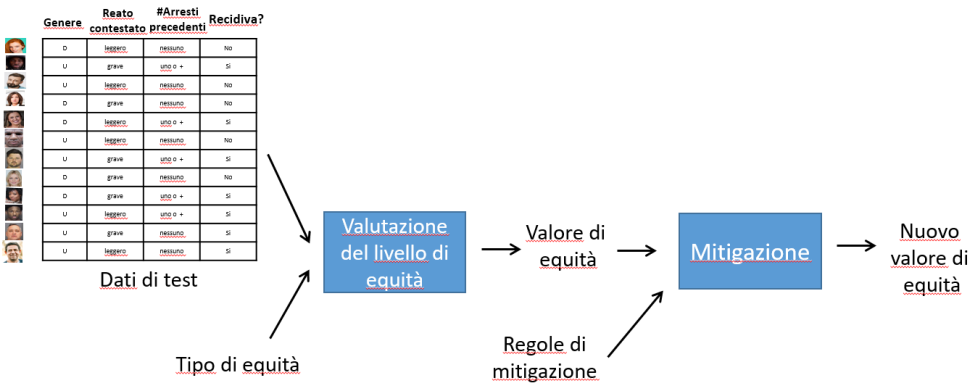


Figura 10.2 – Fasi di valutazione della equità e di mitigazione (della iniquità)

A questo punto vogliamo migliorare la equità, ovvero, detto in altro modo, *mitigare* la iniquità: come si fa? Osserviamo nuovamente il ciclo di vita che ho introdotto nel Capitolo 9, vedi Figura 10.3, in cui ho risistemato i vari blocchi senza alterare il significato.

Nella figura sono incorniciate in rosso le fasi del ciclo di vita di un modello in cui *possiamo effettuare attività di mitigazione*:

- la fase di *rappresentazione/misurazione*, anche chiamata di pre-elaborazione (pre-processing), in cui possiamo modificare i dati che rappresentano il fenomeno di interesse, in modo che il risultato del modello applicato a tali dati sia equo;
- la fase di *modellazione*, anche chiamata di elaborazione (in-processing), in cui viene costruito il modello, viene modificato il modello, o ne viene creato uno nuovo, al fine di migliorare la equità.
- la fase di classificazione, *previsione* e decisione, anche chiamata di post-elaborazione (post-processing), in cui vengono prodotti i dati in output che rappresentano la classificazione o previsione e la relativa decisione finale; in questo caso, per migliorare il livello di equità, le tecniche modificano tale output.

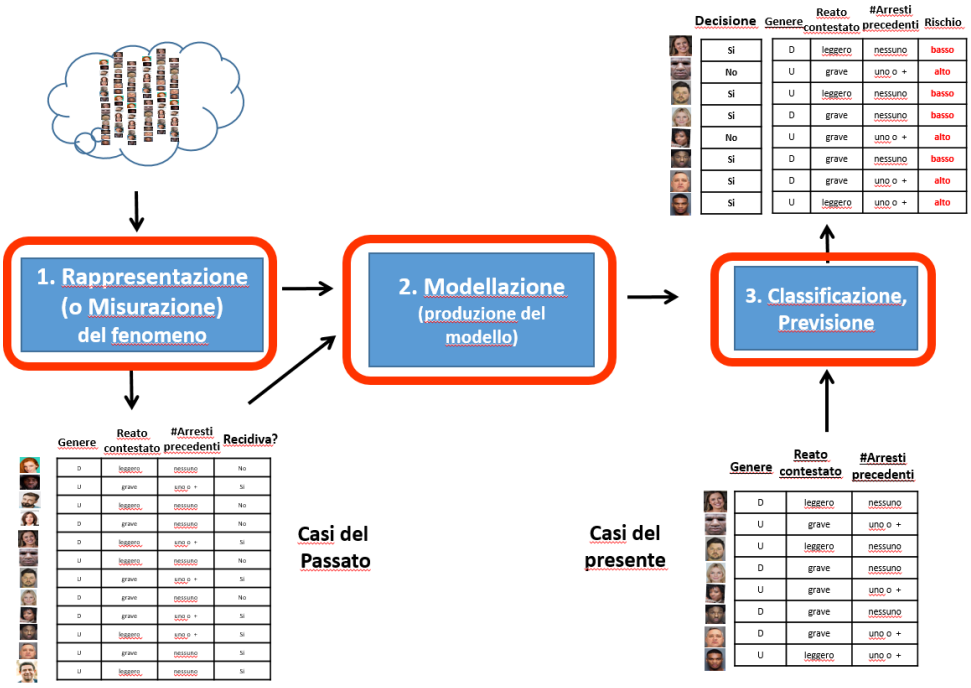


Figura 10.3 – Dove è possibile intervenire sul ciclo di vita del modello per mitigare l'iniquità

E' anche chiaro che le tecniche variano a seconda del tipo di equità oggetto di analisi. Nel seguito del capitolo approfondirò molto la discussione sulla fase di *rappresentazione/ misurazione,* e in qualche misura, le fasi di *modellazione e classificazione.*

Iniziamo!

I dati di ingresso a un modello predittivo sono il modo in cui noi rappresentiamo il mondo. In Figura 10.4 mostriamo i dati di ingresso al modello per il caso della recidiva; l'insieme dei detenuti che possono essere presi in considerazione è chiamato *popolazione*, l'insieme che viene scelto è chiamato *campione*, ogni singolo detenuto è chiamato *elemento*.

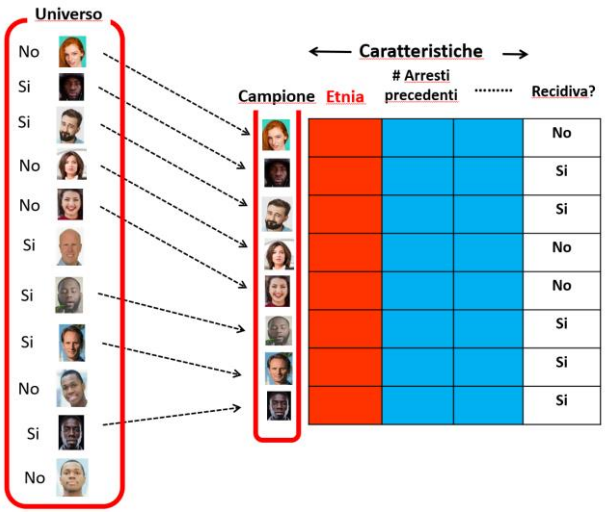


Figura 10.4 – La popolazione rappresentata con un campione di otto elementi e alcune loro caratteristiche

Supponiamo che un modello predittivo verificato sui dati di test abbia rivelato, per esempio, bassi valori della *percentuale di falsi positivi* tra Bianchi e Afro-Americani.

Guardando la Figura 10.4 notiamo subito due cose che ci possono creare sospetti sul *perché* il modello predittivo crei discriminazione:

1. anzitutto, tra le caratteristiche del modello c'è la *etnia*. Non potrebbe essere che questa *caratteristica* sia responsabile della iniquità?

2. un secondo sospetto che ci viene riguarda il passaggio dall'universo al campione; chi ci assicura che a creare iniquità non sia stata la scelta dei detenuti nel campione? In effetti, se facciamo un po' di conteggi, passando da undici elementi (l'universo) a otto elementi (il campione), vengono rappresentati *in proporzione* nel campione molti più Afro-Americani che hanno commesso recidiva rispetto ai Bianchi, vedi Figura 10.5.

	Si	No
B	3	3
AA	3	2

Universo

→

	Si	No
B	2	3
AA	3	0

Campione

Figura 10.5 – Il campione è distorto rispetto all'universo

In questo esempio semplice i due insiemi, l'universo e il campione, hanno un numero di elementi simili, undici elementi e otto elementi; nella realtà il campione può essere, ed è in genere, molto più piccolo dell'universo, vedi Figura 10.6. Questo perché, ad esempio, può accadere che i dati disponibili riguardino solo un numero limitato dei detenuti del passato.

In questo caso, la *scelta degli elementi da includere nel campione* può portare a distorsioni. Se, per esempio, rappresentiamo molti più Bianchi rispetto agli Afro-Americani, il modello predittivo sarà un "vestito" che si adatta molto meglio ai Bianchi che agli Afro-Americani. Ovvero, se tra gli Afro-Americani scegliamo in prevalenza, come in Figura 10.5, quelli che hanno commesso recidiva, il modello formulerà previsioni "sbilanciate", penalizzando gli Afro-Americani.

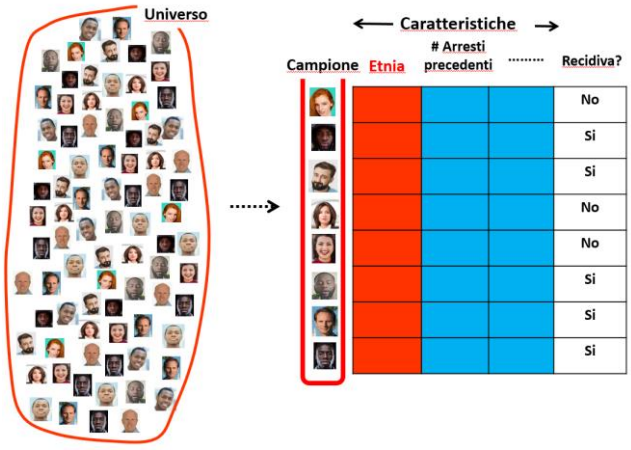


Figura 10.6 – Il campione è usualmente più piccolo dell'universo

Allo stesso modo, anche la *scelta delle caratteristiche* è cruciale nella costruzione del modello predittivo. Per esempio, è possibile che detenuti in penitenziari che forniscono servizi educativi saranno meno portati a commettere recidiva rispetto a detenuti in penitenziari con condizioni di detenzione rigide e non umane. Se non rappresentiamo questa informazione, perderemo per sempre la possibilità di comprendere le cause della recidiva e esprimere questi aspetti nel modello di valutazione del rischio.

Abbiamo dunque visto, con riferimento ai dati di ingresso, che, se vogliamo mitigare la iniquità, potremmo agire sulle caratteristiche (colonne della tabella) ovvero sugli elementi del campione (righe della tabella), vedi Figura 10.7.

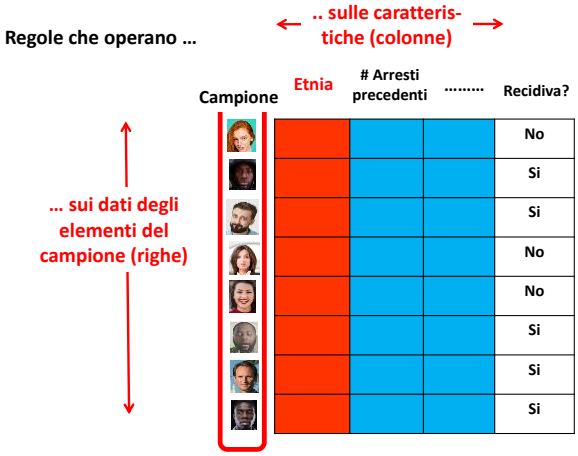


Figura 10.7 – I due tipi di regole che operano sui dati di ingresso

Adesso, per cercare di essere più chiaro, affronterò prima le regole che operano sulle caratteristiche e poi le regole che operano sugli elementi del campione.

<u>Regole di mitigazione che operano sulle caratteristiche</u>

Guardiamo la Figura 10.7, e focalizziamoci sulle caratteristiche *etnia* e *numero di arresti*. Tra le due, l'*etnia* riguarda una proprietà delle persone considerata come *sensibile*. Nel Regolamento per la protezione dei dati personali (GDPR è l'acronimo in inglese), il regolamento che la Unione Europea si è data nel 2018, i *dati sensibili* vengono definiti come i dati che descrivono le caratteristiche delle persone idonee a rivelare l'origine razziale ed etnica, le convinzioni filosofiche, religiose o di altro genere, le opinioni politiche, l'adesione a sindacati, partiti, associazioni od organizzazioni a carattere filosofico, religioso, politico o sindacale, nonché i dati personali idonei a rivelare lo stato di salute e la vita sessuale.

Il GDPR definisce diverse norme per proteggere le persone da un utilizzo distorto dei dati sensibili: per esempio, se il medico di base, anche per mezzo di esami di laboratorio, diagnostica una malattia è tenuto al segreto professionale, e può trasmettere questa informazione solo a soggetti che possono essere interessati in virtù dei servizi che rendono al

malato, come ad esempio altri medici o un ospedale dove la persona viene ricoverata, non ad altri.

Molti dati sensibili, come abbiamo visto nel Capitolo 5, definiscono categorie protette: pensiamo al caso di Rosa Parks e alle discriminazioni sui posti in autobus e tanti altri servizi sociali, o alla tragedia dell'Olocausto perpetrata dalla Germania nazista verso gli ebrei, oltre che verso altre etnie. Insomma, per tante ragioni, possiamo arrivare alla conclusione che un primo metodo per ripristinare la equità sia (ricordo che Regola T significa regola tecnologica):

Regola T1 - Escludere dalle caratteristiche quelle protette (fairness as blindness, equità come cecità)

Questa prima regola ci ricorda una delle immagini con cui spesso è raffigurata la Giustizia, quella di Figura 10.8, la Giustizia bendata.

Figura 10.8 – La giustizia bendata

Accade in alcune procedure decisionali, per esempio un concorso, che, per non essere influenzati nelle valutazioni, non si voglia avere nessuna informazione sulle persone partecipanti; nelle valutazioni dei lavori di ricerca da accettare in alcune conferenze o riviste questo criterio porta a cancellare i nomi e cognomi degli autori dei lavori da mandare in valutazione.

Nell'ambito delle equità introdotte nel Capitolo 8, escludere le caratteristiche protette compare tra le equità basate sulla *ricerca delle cause*, vedi Figura 10.9: siccome io sospetto che la equità sia violata da una caratteristica protetta, la escludo, e così non ci penso più.

Tipi di equità (fairness) basate su…	**Definizione**
…. ragionamento causale (perchè?) **1. Equità ottenuta attraverso la non consapevolezza** 2. Equità ottenuta attraverso la consapevolezza	1. Nessun attributo sensibile è utilizzato nel processo predittivo 2. Elementi simili secondo una funzione di distanza devono avere valori simili nella predizione

Figura 10.9 – Equità ottenuta attraverso la non consapevolezza

Bene, proviamo ad applicare questo metodo al caso della recidiva; partendo dalle due caratteristiche *etnia* e *numero di arresti*, vediamo ora un esercizio così concepito. Vogliamo valutare le due misure di equità:

- Percentuale di falsi positivi e

- Percentuale di falsi negativi

su dati di input definiti inizialmente su *etnia* e *numero di arresti*, e successivamente solo sul *numero di arresti*, vedi Figura 10.10.

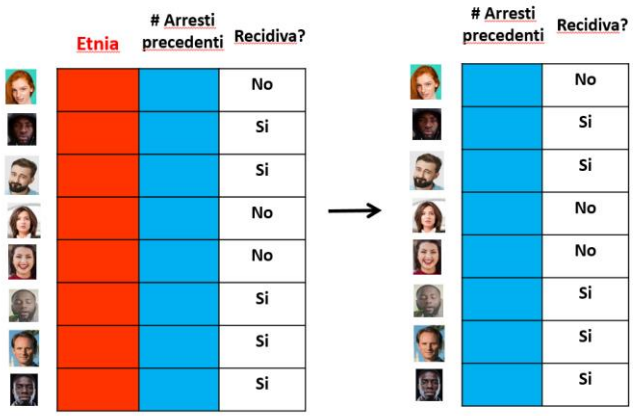


Figura 10.10 – Applichiamo la Regola T1: escludere dalle caratteristiche l'etnia

L'esercizio parte dall'albero di decisione A di Figura 10.11.

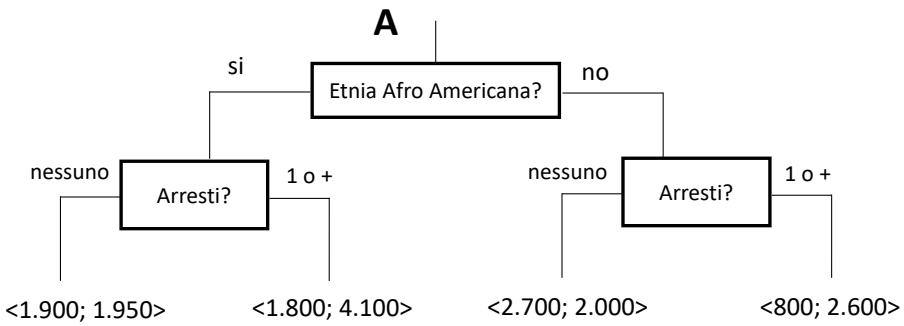


Figura 10.11 – L'albero di decisione A con etnia e numero di arresti

Esercizio 10.1

1. Produrre il modello predittivo per l'albero di Figura 10.11;
2. Calcolare il numero di veri e falsi positivi e negativi per le due etnie;
3. Calcolare le due misure di equità costituite da

a. Percentuale di Falsi positivi /(Falsi positivi + Veri negativi)

b. Percentuale di Falsi Negativi /(Falsi negativi + Veri positivi)

(che corrispondono alle due equità di ProPublica).

Le risposte nella prossima pagina.

Ecco le soluzioni all'Esercizio 10.1.

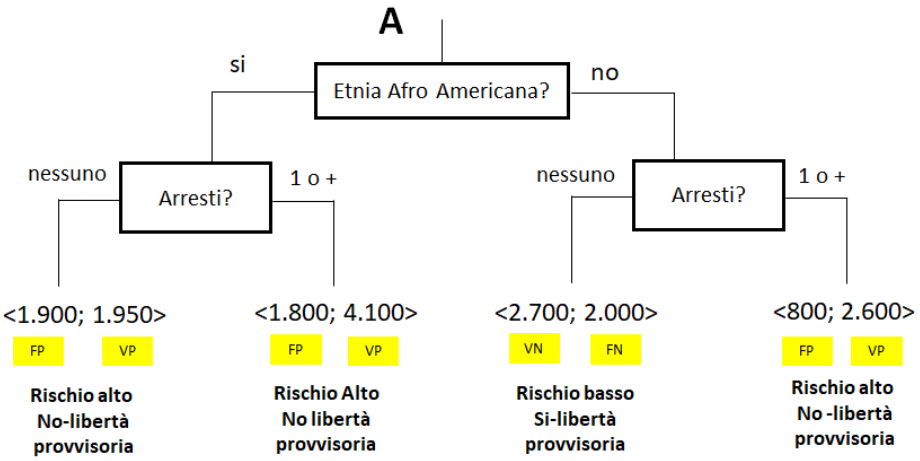


Figura 10.12 – Il modello predittivo, i veri e falsi positivi e negativi

A

Etnia	Tipologia Universo	Numerosità
Afro Americani	Veri Positivi	1.950 + 4.100
Afro Americani	Falsi Positivi	1.900 + 1.800
Afro Americani	Veri Negativi	0
Afro Americani	Falsi Negativi	0
Bianchi	Veri Positivi	2.600
Bianchi	Falsi Positivi	800
Bianchi	Veri Negativi	2.700
Bianchi	Falsi Negativi	2.000

	FP/FP+VN	FN/FN+VP
Afro Americani	0.23	1
Bianchi	0.43	0

Figura 10.13 – Le due misure di equità

A questo punto svolgiamo il seguente esercizio.

Esercizio 10.2 - Adesso vi propongo di:

1. Produrre il nuovo albero di decisione B associato all'unica caratteristica *Numero degli arresti.*
2. Produrre il nuovo modello predittivo
3. Calcolare il nuovo numero di veri e falsi positivi e negativi per le due etnie
3. Calcolare le nuove misure di equità costituite da
a. Percentuale di Falsi positivi: Falsi positivi / (Falsi positivi + Veri negativi)
b. Percentuale di Falsi negativi: Falsi negativi / (Falsi negativi + Veri positivi)

Ecco le soluzioni all'Esercizio 10.2.

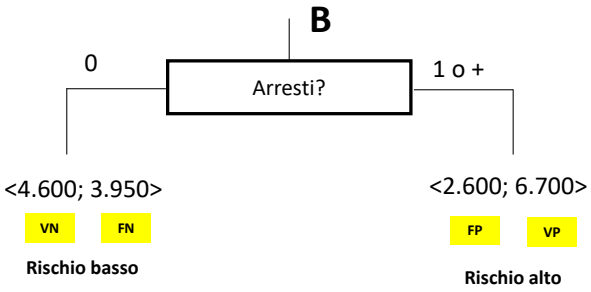


Figura 10.14 – Nuovo albero di decisione B, modello predittivo, veri e falsi positivi e negativi

A

Etnia	Tipologia Universo	Numerosità
Afro Americani	Veri Positivi	1.950 + 4.100
Afro Americani	Falsi Positivi	1.900 + 1.800
Afro Americani	Veri Negativi	0
Afro Americani	Falsi Negativi	0
Bianchi	Veri Positivi	2.600
Bianchi	Falsi Positivi	800
Bianchi	Veri Negativi	2.700
Bianchi	Falsi Negativi	2.000

B

Etnia	Tipologia Universo	Numerosità
Afro Americani	Veri Positivi	4.100
Afro Americani	Falsi Positivi	1.800
Afro Americani	Veri Negativi	1.900
Afro Americani	Falsi Negativi	1.950
Bianchi	Veri Positivi	2.600
Bianchi	Falsi Positivi	800
Bianchi	Veri Negativi	2.700
Bianchi	Falsi Negativi	2.000

Figura 10.15 – Confronto tra veri e falsi positivi e negativi per i modelli A e B

A	FP/FP+VN	FN/FN+VP
Afro Americani	1	n.d.
Bianchi	0.23	0.43

B	FP/FP+VN	FN/FN+VP
Afro Americani	0.49	0,31
Bianchi	0.23	0,43

Figura 10.16 – Confronto tra misure di equità per i modelli A e B.

Dunque, in questo caso l'eliminazione della etnia dal modello predittivo ha migliorato le due "equità ProPublica", perché gli Afro-Americani avevano una distribuzione dei casi sfavorevole rispetto a quella dei Bianchi (due valori molto vicini per la foglia più a sinistra dell'albero), e quindi eliminare la caratteristica *etnia* ha portato a un albero che ignora questo aspetto.

Attenzione però: l'eliminazione dal modello della caratteristica sensibile *può non bastare a rimuovere la discriminazione.* Spesso, infatti, tra le caratteristiche se ne annidano altre che *non sono indipendenti dalla caratteristica sensibile*, queste caratteristiche sono dette *proxi*.

Facciamo un'esempio. Uno storico strumento di discriminazione utilizzato negli Stati Uniti per discriminare determinate etnie fu il cosiddetto *redlining*, letteralmente *tracciare una linea rossa*, la pratica, cioè, di escludere o limitare determinate popolazioni nella possibilità di accedere a servizi sociali.

Non sto dicendo niente di nuovo! Rosa Parks e tutti gli Afro-Americani non potevano sedersi nei posti davanti negli autobus, vedi in Figura 10.17 lo storico autobus dove Rosa Parks fece il gesto di disobbedienza civile di cui abbiamo parlato nel Capitolo 5.

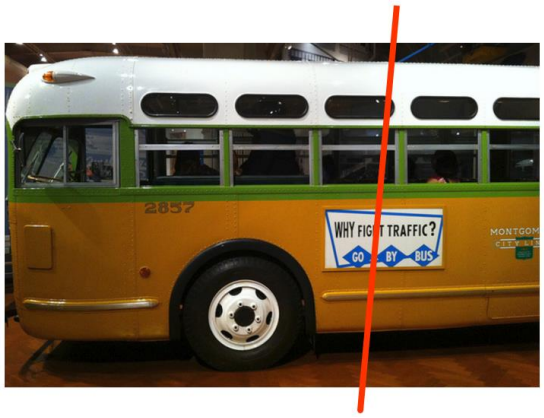


Figura 10.17 – L'autobus di Rosa Parks e il redlining

Un altro tipo di servizio sociale a lungo negato a particolari etnie, principalmente gli Afro-Americani, sono stati i prestiti bancari. La discriminazione avveniva *colorando di rosso* determinati quartieri abitati prevalentemente da Afro-Americani, vedi Figura 10.18; fino agli anni 90 del secolo scorso, molte banche prestavano denaro a Bianchi a basso reddito, e lo negavano a Afro-Americani a medio o alto reddito.

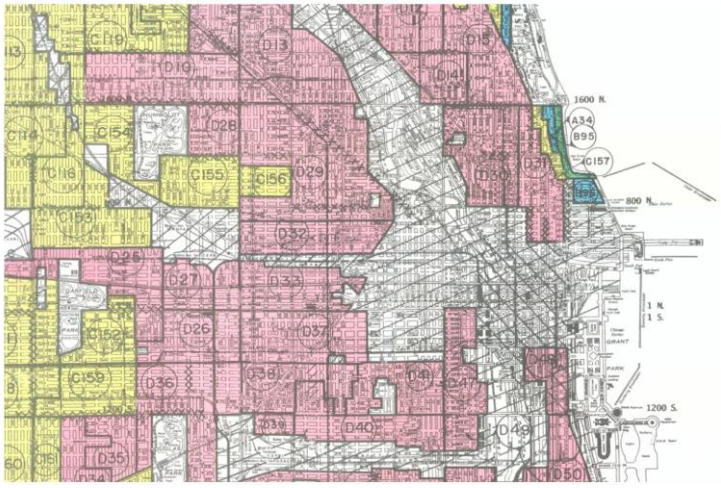


Figura 10.18 – Il fenomeno del redlining nella concessione di prestiti

In pratica, il quartiere in cui abitava il richiedente prestito veniva visto come una informazione approssimata (proxi) per identificare la etnia. Tale profilazione portava a profitti maggiori per le banche; in ogni caso, era *eticamente e legalmente inaccettabile*.

I fenomeni di redlining sono di fatto *discriminazioni indirette*, basate su caratteristiche, nel nostro caso il quartiere, che non rappresrntano la etnia, *ma che con la etnia sono fortemente correlate.*

Un'altra caratteristica tipicamente correlata con le condizioni economiche e con le caratteristiche protette è il codice di avviamento postale (CAP); negli Stati Uniti lo zip code (il nostro CAP) è fortemente correlato con la etnia. Se non è ammesso usare la etnia nelle decisioni, non deve essere ammesso neanche usare variabili proxi.

Scusa cosa intendi per correlato?

La *correlazione* in statistica è una relazione tra due caratteristiche, tale che a ciascun valore della prima corrisponda un valore della seconda, *seguendo una certa regolarità*. La correlazione non dipende da un rapporto di causa-effetto, quanto dalla tendenza di una caratteristica di cambiare in funzione di un'altra.

La *correlazione positiva*, per cui all'aumentare del valore di una caratteristica aumenta il valore dell'altra, può assumere valori nell'intervallo [0,1]; 0 significa nessuna correlazione, 1 significa che il legame tra le due caratteristiche è espresso da una precisa funzione, cioè le due caratteristiche crescono nello stesso modo.

Ad esempio nel lavoro [19] viene citato l'esempio di un insieme di dati chiamato German Credit, che riporta le decisioni connesse alla concessione di prestiti bancari. Le decisioni sono correlate con l'età dei richiedenti il prestito, e la correlazione ha valore 0.09. Supponiamo che usare l'età sia proibito per legge; se rimuoviamo *l'età* dalle caratteristiche dei dati, ciò non rimuoverà la discriminazione basata sulla età, perché altre caratteristiche, come il possesso di una casa, forniscono informazioni indirette sulla età del richiedente il prestito; infatti la caratteristica è correlata con età con un un valore superiore a 0.1.

Un altro esempio, tratto dallo stesso lavoro, riguarda una azienda di consulenza aziendale, accusata in una causa di discriminazione verso minoranze etniche. L'azienda aveva usato informazioni sui *reati pregressi* per escludere candidati alla assunzione durante lo screening iniziale che precedeva i colloqui. Nei dati disponibili non era compresa la etnia, ma, in virtù della correlazione tra reati pregressi e etnia, di fatto il solo considerare i reati pregressi risultava discriminatorio. La corte condannò l'azienda, perché arrivò alla conclusione che l'aver compiuto reati *non poteva essere considerato rilevante per l'assunzione*.

Insomma, anche quando cerchiamo di rimuovere le caratteristiche protette, abbiamo tanti strumenti evidenti e nascosti per inferire tali caratteristiche; e queste inferenze *possono essere meno accurate degli stessi dati che abbiamo cancellato*. Ad esempio, sono riportati casi negli Stati Uniti, in cui rimuovere la etnia ha *acuito e non migliorato* le discriminazioni nell'impiego degli Afro-Americani.

.
Gli esempi precedenti ci forniscono la seguente regola

[19] A Kamiran et al. - Quantifying explainable discrimination and removing illegal discrimination in automated decision making, Knowledge Information Systems (2013) 35:613–644

Regola T2 – Accanto alla caratteristica protetta, verificare anche l'incidenza sulla iniquità delle caratteristiche ad esse correlate (caratteristiche proxi)

Esiste un altro problema che ci deve far essere cauti con la eliminazione delle caratteristiche protette, o loro proxi, ed è la *relazione con la accuratezza*. Si rischia di eliminare le caratteristiche con alto guadagno informativo, e di basarsi su caratteristiche che non riescono a discriminare sufficientemente i gruppi di elementi nelle foglie, dando luogo a un alto numero di falsi positivi e falsi negativi. Per cui vale la regola:

Regola T3 – Fare attenzione a che l'eliminazione di caratteristiche protette o proxi non peggiori in modo inaccettabile la accuratezza.

E poi, non sempre occultare la etnia nei modelli è una buona idea. Diversi lavori nella ricerca clinica e farmacologica[20] propongono farmaci *per specifiche etnie*. Ad esempio il BiDil, usato per il trattamento della insufficienza cardiaca come complemento alla terapia tradizionale, è indicato per pazienti che si sono identificati come Afro-Americani.

Inoltre, ci possono essere casi in cui il dato protetto è quello che ha maggiore *guadagno informativo*. Proprio nei modelli per la recidiva, il genere è un'importante predittore, le donne recidivano meno degli uomini, così che eliminando il genere nel modello dà luogo a *maggiore, e non minore,* discriminazione verso le donne. Per cui possiamo formulare la regola:

Regola T4 – Eliminare la caratteristica protetta, o caratteristiche proxi ad essa correlate, non sempre è una buona idea, perché può ridurre la accuratezza del modello, e quindi penalizzare le popolazioni che si vuole proteggere.

Introduco ora una quinta regola; per farlo ho bisogno di un esempio introduttivo.

Il caso che consideriamo nel seguito riguarda gli esami di ammissione a una università che ha due corsi di laurea, medicina e informatica. Tutti i partecipanti alla selezione fanno un test, la decisione di ammettere o meno il partecipante è fatta individualmente per ogni partecipante a seguito di una intervista. In Figura 10.19 riportiamo un primo esempio di partecipanti e di risultati della ammissione, per i due corsi di laurea medicina e informatica.

Lo studio di caso mostrerà che si possono distinguere tra due tipi di discriminazione, quella che può essere giustificata sulla base, ad esempio, di una differente numerosità degli uomini e donne coinvolte nel modello classificatorio, *discriminazione* che chiameremo *spiegabile*, e quella che non può essere giustificata, che chiameremo discriminazione *illegale*.

Una *discriminazione spiegabile* lo è sulla base di una o più caratteristiche (nell'esempio che faremo, il corso di laurea cui fanno domanda partecipanti alla selezione), tali caratteristiche assumono il nome di *caratteristiche esplicative*.

[20] Vedi ad esempio Vence L. Bonham, J.D., Shawneequa L. Callier, J.D., and Charmaine D. Royal, Will Precision Medicine Move Us beyond Race? The New England Journal of Medicine, 2016.

	Medicina		Informatica	
	Donne	Uomini	Donne	Uomini
Numero di domande	800	200	200	800
Tasso di accettazione (%)	20	20	40	40
Ammessi	160	40	80	320

Figura 10.19 – Un possibile esempio di partecipanti e risultati della ammissione

Una *caratteristica esplicativa* (ad esempio il corso di laurea) è la caratteristica, tra le tante, che viene usata per comprendere se la discriminazione sia spiegabile senza invalidare la equità; sia spiegabile, cioè, unicamente sulla base della informazione obiettiva portata dalla caratteristica. Essa può essere misurata, nel nostro esempio, come il guadagno informativo:

1. sul genere, dato il corso di laurea cui è stata fatta domanda di ammissione, ovvero
2. sulla ammissione, dato il corso di laurea in cui è stata fatta domanda.

La caratteristica esplicativa va scelta tra quelle che non sono correlate con la caratteristica sensibile. Per esempio, quando il genere è la caratteristica sensibile, la scelta della relazione di parentela come marito o moglie non è una buona scelta come *caratteristica esplicativa*, perché *genere* e *relazione di parentela* sono molto correlate come significato. Al contrario, le ore lavorate in una settimana sono una buona caratteristica esplicativa per la retribuzione.

Assumiamo nel seguito come criterio di equità la *parità demografica*.

Esempio 1 – Nei dati di ingresso esiste solo discriminazione in termini di parità demografica, spiegabile. In questo esempio tutta la discriminazione è spiegabile, e non vi è discriminazione "illegale".

L'esempio è quello della Figura 10.20, che estende la Figura 10.19; supponiamo che a un concorso di ammissione a due corsi di laurea di una università, medicina e informatica, si presentino 2.000 candidati, di cui 1.000 uomini e 1.000 donne. Ognuno dei due corsi riceve lo stesso numero di domande, ma medicina è più richiesta tra le donne, da cui provengono l'80% delle domande. Assumiamo che medicina sia più competitiva, cioè il numero di posti sia inferiore a quello di informatica.

All'interno di ognuno dei due corsi di laurea, donne e uomini sono trattati allo stesso modo, come mostrato in Figura 10.20. Tuttavia, le percentuali aggregate sul totale degli ammessi indica che sono stati ammessi il il 24% delle donne e il 36% degli uomini, a fronte di un tasso di accettazione del 20% in medicina e 20% in informatica. La differenza è spiegata (o giustificata, se vi piace di più…) dal fatto che più donne hanno fatto domanda per il corso di laurea più competitivo, cioè con meno posti.

	Medicina		Informatica	
	Donne	Uomini	Donne	Uomini
Numero di domande	800	200	200	800
Tasso di accettazione (%)	20	20	40	40
Ammessi	160	40	80	320
Donne (%)	24			
Uomini (%)	36			

Figura 10.20 – Caso di differenza percentuale di ammissione completamente spiegabile

Dalla Figura 10.20 noi concludiamo che la differenza percentuale spiegabile tra tassi di accettazione di donne e uomini è D%spiegabile = 36% - 24% = 12%.

Nell'esempio 2 occorrono sia casi di discriminazione spiegabile sia casi di discriminazione illegale. In questo caso, vedi Figura 10.21, le percentuali degli ammessi tra uomini e donne sono del 19% e del 41%, a fronte di un tasso di accettazione complessivo del 17% in medicina e del 43% in informatica. Il nostro scopo, ora, è determinare quale parte della differenza tra donne e uomini è spiegabile dalle diverse percentuali di partecipanti a medicina e a informatica, e quale a una discriminazione che non rispetta la parità demografica.

	Medicina		Informatica	
	Donne	Uomini	Donne	Uomini
Numero di domande	800	200	200	800
Tasso di accettazione (%)	15	25	35	45
Ammessi	120	50	70	360
Donne (%)	19			
Uomini (%)	41			

Figura 10.21 – Caso di differenza percentuale di ammissione completamente spiegabile

Dalla Figura 10.21 vediamo che, nel secondo caso la differenza percentuale tra tassi di accettazione di donne e uomini è D%totale (cioè la somma tra spiegabile e contraria alla parità demografica) = 41% - 19% = 22%. La conclusione è che la differenza percentuale inaccettabile rispetto al criterio di parità demografica = 22% - 12% = 10%.

I precedenti esempi ci portano alla seguente regola:

Regola T5 – Scelte la caratteristica esplicativa e la caratteristica protetta, prima di mitigare la iniquità separare la discriminazione spiegabile da quella non giustificabile sulla base del criterio di equità scelto.

Fino ad ora abbiamo visto come separare le discriminazioni spiegabili da quelle illegali. Come facciamo ora a ristabilire la equità violata dalla discriminazione illegale, ad esempio nel caso di Figura 10.21?

Nell'esempio precedente, le donne siano state discriminate nell'accesso alla università del 10%, La mitigazione avviene in questo modo; prima di tutto, ordiniamo donne e uomini secondo un criterio che definisca una precisa graduatoria; per esempio, sulla base del voto ottenuto al termine dell'esame.

Si, ma nel caso dei detenuti la recidiva assume solo due valori, si e no!

E' vero, accipicchia come ricordi bene! Possiamo ordinarli sulla base del numero degli arresti pregressi: 0, 1, 2, 3, ecc. poi, per tutti i detenuti con lo stesso numero di arresti, assumendo di conoscere la data di nascita in ordine di età.

A questo punto, possiamo individuare un certo numero di donne che sono collocate immediatamente sotto la soglia della accettazione, *collocandole al di sopra*, e un uguale numero di uomini che sono collocati immediatamente al di sopra, *collocandoli al di sotto*.

Cosa significa un certo numero?

Giusta domanda. Significa tanti quanti sono necessari per portare a 0 quel 10% di discriminazione illegale… vedi Figura 10.22.

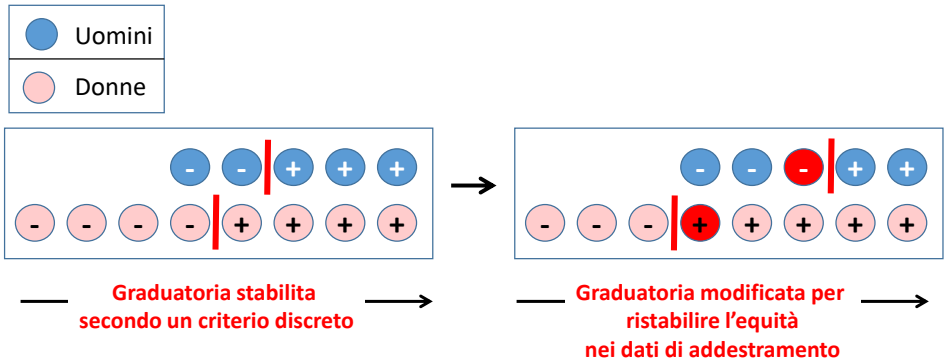


Figura 10.22 – Come opera la tecnica di mitigazione sui dati di addestramento

Sulla base dell'esempio possiamo introdurre la seguente regola per mitigare la discriminazione.

Regola T6 – Quando vi sia una discriminazione non spiegabile che viola la equità considerata, ristabilire la equità intervenendo sui dati di addestramento.

Notate che in questo caso abbiamo una *tensione* tra equità di gruppo (in questo caso equità di genere) e equità individuale; per ristabilire la equità di gruppo tra uomini e donne, siamo costretti a *violare la equità individuale tra gli uomini che abbiamo retrocesso e le donne che abbiamo promosso*.

Regole di mitigazione sui dati del campione

Qui abbiamo una sola regola:

Regola T7 – Rendere rappresentativo il campione

Abbiamo visto nel primo esempio del capitolo (Figura 10.4) che il campione scelto non rappresentava in modo equilibrato l'universo di partenza, e gli Afro-Americani era scelti molto più dei Bianchi tra i soggetti colpevoli di recidiva.

Dobbiamo quindi costruire il campione rappresentando *tutti i gruppi definiti dalla caratteristica sensibile in modo statisticamente omogeneo*. Ad esempio, nelle figure 10.23 e 10.24 mostriamo come, cancellando due elementi dell'universo e sostituendoli con due nuovi elementi, possiamo equalizzare i Bianchi e gli Afro-Americani.

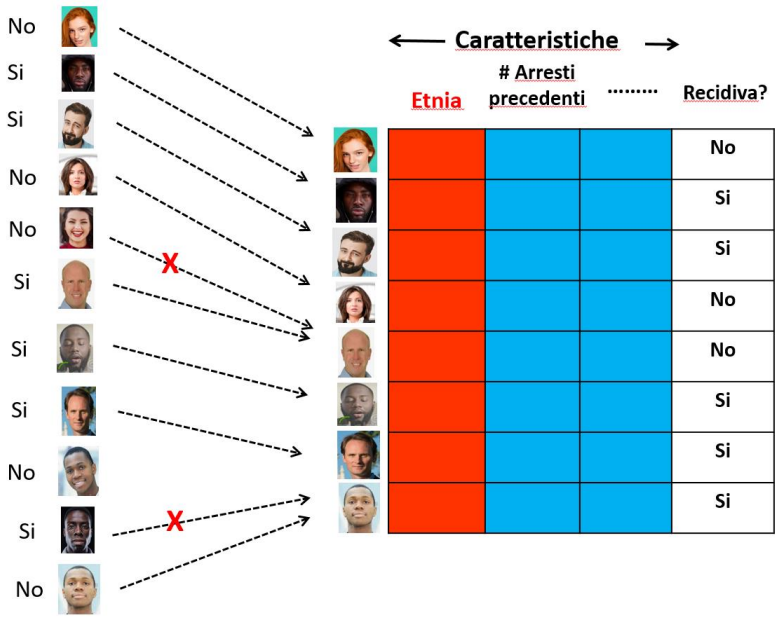


Figura 10.23 – Omogeneizzare il campione

	Si	No
B	3	3
AA	3	2

→

	Si	No
B	3	3
AA	3	2

Figura 10.24 – Nuova suddivisione dei Bianchi e Afro-Americani

Regole di mitigazione su modello e tecnica

Riguardo al modello e alla tecnica di Machine learning usata per costruirlo, abbiamo visto ad esempio che le tecniche basate su foreste causali sono più accurate delle tecniche basate su

alberi di decisione. Avere una buona accuratezza per il modello è un obiettivo importante, perchè modelli inaccurati sono un po' come .,,.un terno al lotto, non ci possiamo fidare molto. Una regola che possiamo applicare in questo caso è la seguente:

Regola T8 – Quando il modello è inaccurato, cambia la tecnica usata per generarlo, così da accrescerne la accuratezza.

Regole sui dati in output

La regola in questo caso è la seguente:

Regola T9 – Modifica i dati prodotti in output dal modello così da ristabilire la equità

La regola di mitigazione T9 ci dice che per costruire l'equità possiamo modificare i *risultati* prodotti da un modello classificatorio. Gli approcci di post-elaborazione partono dal riconoscimento che l'output di un modello predittivo può essere iniquo verso una o più caratteristiche; applicano trasformazioni sull'output che migliorano la equità, senza dover operare sui dati di ingresso o sul modello. In questo modo. sono applicabili praticamente sempre, anche quando vi sia un segreto industriale sul modello, ovvero chi lo ha prodotto non intenda svelare i dati di addestramento.

Chiaramente, questi approcci sono i più facili da usare, perché modificano le decisioni, non il processo o i dati di addestramento che le hanno generate. Per evitare abusi, anche in questo caso occorre applicare un metodo trasparente e razionale nella modifica dell'output. ad esempio, il metodo che abbiamo visto applicato sui dati di addestramento nel caso della regola T6, in cui la graduatoria tra i partecipanti veniva modificata sugli elementi del campione vicini alla soglia di ammissione alla Università, può essere applicato sulla graduatoria in output.

Attenzione però, perché anche in questo caso cercando di mitigare la iniquità verso gruppi, si viola la equità verso individui. Vediamo anche qui un esempio di tradeoff.

Per concludere questa sezione sulla regole T di mitigazione della iniquità tramite interventi su dati e modelli, vediamo in Figura 10.25 il quadro generale comune a tutte i tipi di mitigazione descritti.

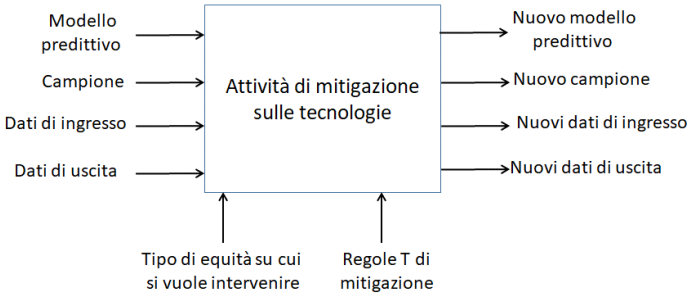


Figura 10.25 – Elementi considerati nelle regole di mitigazione T su modello e dati

10.3 Regole di mitigazione sociali (Regole S)

Queste regole riguardano le responsabilità della società e di tutti noi nell'uso dei modelli classificatori e predittivi. Sono regole che non dipendono dal modello e dai dati, *dipendono da tutti noi*. Le enuniciamo in termini generali, ma è chiaro che volta a volta i protagonisti nella applicazione di queste regole sono *i singoli, le comunità, i gruppi sociali, il Parlamento, il Governo*. Faremo in questo senso alcuni esempi.

Allacciamo le cinture, vengono affrontate questioni di base della nostra vita!

Regola S1 – Qualora l'uso delle caratteristiche protette nei modelli non sia regolamentato per legge, valutare se sia opportuno intraprendere una iniziativa legislativa

Abbiamo già visto come nel passato alcune pratiche, come il redlining per i prestiti bancari, siano state a un certo punto proibilte per legge. Questo significa che da un certo momento in poi, se in una banca veniva utilizzato un modello predittivo per prendere decisioni, tale modello *non poteva* inglobare il redlining tra le caratteristiche di ingresso al modello.

Ogniqualvolta in una società si avverte l'uso di pratiche analoghe al redlining, le comunità e i movimenti di opinione possono attivarsi per sensibilizzare il Parlamento e il Governo a intraprendere iniziative legislative pe rimuovere tali pratiche dai modelli adottati per prendere decisioni.

Regola S2 – Quando viene usato un modello classificatorio o predittivo, definisci per legge una soglia di iniquità considerata accettabile che, se viene superata, può giustificare l'intervento di una terza parte giuridica che risolve il conflitto

Rispondono a questa regola le Linea guida formulate negli Stati Uniti dalla Commissione Federale sulle Pari opportunità di impiego: nei concorsi di assunzione in enti pubblici o ammissione alla Università, se la *parità demografica* è violata rispetto all'effettiva percentuale dei positivi per più del 20%, ciò costituisce una violazione della equità che giustifica una causa legale. Peraltro, la violazione non è di per sè illegale, per individuare una responsabilità la disparità deve essere ingiustificabile o evitabile.

Occorre avvertire che le linee guida, pur se espresse in forma quantitativa (il 20% in più) non forniscono risposte semplici e definite in modo inequivocabile. Inoltre, se garantiscono una equità di gruppo, portano a violare la equità tra individui caratterizzati dalle stesse caratteristiche appartenenti a gruppi sociali differenti.

Regola S3 – Accanto alla equità comparativa, persegui la equità non comparativa o assoluta

Per tutto il libro ho discusso equità *comparative*, equità, cioè, che confrontano gruppi sociali o individui e *determinano tramite confronti* se i risultati dei modelli predittivi creino o non creino discriminazioni tra i gruppi e individui.

C'è nelle nostre vite un'altra forma di equità, di natura *non comparativa o assoluta*. E' l'equità per cui a un gruppo sociale o a un individuo debba essere garantito una determinata risorsa, servizio, diritto, *indipendentemente da un confronto con altri gruppi sociali, per il solo fatto, cioè, che quel gruppo sociale o individuo esiste e opera nella società.*

In altre parole, interventi che combattono ingiustizie sociali, possono aumentare la disparità di trattamento, nell'ottica di perseguire oltre che una equità comparativa, un secondo tipo di equità, che chiamiamo *assoluta*, e che obbliga a *trattare ogni soggetto come dovrebbe essere trattato, indipendentemente da come vengono trattati tutti gli altri*.

La regola S4 è espressa da una domanda.

Regola S4 – Per ristabilire la equità, dobbiamo essere certe volte iniqui, e adottare una equità riparativa?

Questa regola sorge in casi come i seguenti: esistono differenze rispetto agli uomini riguardo alla riproduzione e alle responsabilità conferite dalla società alle donne *per la educazione dei figli*. Queste differenze confliggono con i percorsi e gli avanzamenti di carriera, avanzamenti, ad esempio, delle ricercatrici nell'ottenere lo statuto di professore o delle avvocatesse nell'ottenere una modifica del proprio status da apprendiste a associate in uno studio legale.

Se le prestazioni lavorative di queste donne sono giudicate inferiori confrontate con le prestazioni degli avvocati maschi, la cui carriere non è frenata da responsabilità nella educazione dei figli, allora è necessario dare, ad esempio, alle ricercatrici e alle avvocatesse, un trattamento speciale; si deve applicare nel loro caso una *equità riparatoria*, che viola la supposta oggettività dei criteri che si ispirano alla *esclusione delle caratteristiche protette* (Regola T1) come li abbiamo discussi in precedenza.

Il tentativo di rendere nitido e oggettivo il criterio della equità attraverso la esclusione delle caratteristiche protette, ci porta al rischio di percorre due cammini, entrambi, con una metafora da montanaro, *accidentati*. O ignoriamo le differenze e le discriminazioni che occorrono nella realtà, a favore di un impegno ideologico a essere ciechi e agnostici, ovvero assumiamo un argomento orientato, quando il contesto la richiede, a quella che ho chiamato poco fa "equità riparativa", aprendo il fianco ad essere vulnerabili ad accuse di inconsistenza, di applicare due pesi e due misure. L'equità sembra essere meglio servita da una osservazione sensibile della realtà che da un rifiuto ideologico a guardare la realtà.

Regola S5 – La equità di gruppo e la equità tra individui sono talvolta inconciliabili, e occorre decidere tra l'una e l'altra.

Questa regola, che abbiamo già visto applicata in alcuni esempi, è stata oggetto di molte discussioni e prese di posizione, e si scontrano due scuole di pensiero, a favore dell'una e

dell'altra, vedi ad esempio il libro [21], nettamente schierato a favore della equità di gruppo; ciò sulla base dell'argomento che essa rispetta *un criterio generale valido per l'intera società*.

In paesi con forte tradizione di iniziativa individuale, tende a prevalere il criterio dei diritti dei singoli individui.

Regola S6 – E' spesso impossibile mitigare tutte le iniquità, e scegliere quella o quelle da mitigare è spesso un problema di tensione tra diverse visioni della società

Abbiamo visto questa regola nel caso della controversia Propubllca verso Northpointe; ProPublica mostra come il modello predittivo sia iniquo rispetto a due misure di equità, e Northpointe mostra come sia equo per altre due misure. Chi ha ragione? Secondo me, la questione: chi dei due ha ragione? è malposta; dipende dagli equilibri e dalle tensioni tra le diverse forze che si confrontano nella società, da cui può scaturire storicamente il prevalere dell'una o dell'altra forma di equità.

La legislazione può svolgere il ruolo espresso dalla seconda definizione di obiettività che abbiamo investigato nel Capitolo 5, quella che vede la equità come la capacità di individuare soluzioni equilibrate tra visioni del mondo o teorie *in conflitto tra di loro.* Ma a questo punto la domanda, a cui non ho risposta, diventa: cosa significa equilibrate?

Regola S7 – Combatti sempre affinchè i grandi utilizzatori di dati e modelli migliorino continuamente le loro procedure di mitigazione della iniquità, sei tu l'utente! Le grandi aziende che operano a livello mondiale nel mercato dei dati digitali devono essere trasparenti riguardo ai modelli che usano e devono migliorare continuamente i propri modelli per renderli più equi.

Le grandi aziende di cui parliamo sono naturalmente Google, Amazon, Microsoft e tutte le aziende in generale che fanno profitti nella economia dei dati digitali. Dobbiamo pretendere che:

1. facciano ricerca per migliorare continuamente la equità dei propri modelli e connessi servizi, documentando nel tempo risultati progressivi di miglioramento.
2. realizzino e rendano disponibili siti e corsi per aumentare la consapevolezza e conoscenza del processo di misurazione e mitigazione delle iniquità.
3. creino e mettano a disposizione ambienti aperti per misurare e mitigare la equità.

Riguardo al punto 1, torniamo all'esempio di traduzione in Google Translator dall'inglese all'italiano di *doctor* in dottore e *nurse* in infermiera; Google ha una linea di ricerca per la individuazione e correzione sistematica di distorsioni di genere nelle traduzioni tra lingue naturali. L'idea è di generare traduzioni sia con genere maschile che con genere femminile, riducendo in questo modo la distorsione guidata dal genere. La tecnica include tre fasi:

1. la individuazione della distorsione linguistica,
2. la generazione di alternative e
3. la scelta tra alternative e la validazione del risultato, prima della produzione della traduzione all'utente.

[21] F. Schauer - Profiles, probabilities and stereotypes, Harvard Univeristy Press, 2009.

Questo approccio è stato dapprima usato nelle traduzioni tra turco e Inglese (2018) e viene esteso nel tempo verso altre lingue naturali, assumendo sempre l'Inglese come seconda lingua coinvolta nella traduzione.

Per i punti 2 e 3, vedremo in Appendice 1 alcuni siti realizzati da Google, IBM, Microsoft e Lime.

Regola S8 - Pretendi che i modelli predittivi siano usati in modo trasparente

All'epoca del caso ProPublica & Northpointe, quest'ultima si rifiutò di rivelare il funzionamento della tecnica adottata per produrre il modello predittivo, rendendo impossibile nei fatti comprendere le ragioni e la natura dei risultati prodotti.

Ciò è comprensibile, nel senso che secondo la legislazione vigente Northpointe era autorizzata a non svelare un segreto industriale, e svelarlo avrebbe comportato minore competitività sul mercato. Ma suscita anche interrogativi su quanto in una Società sia possibile che una azienda sviluppi e venda modelli di valutazione del rischio che aiutano a prendere decisioni sulla vita delle persone, senza che la Società abbia la possibilità di *guardare dentro* questi modelli.

Abbiamo dunque la seguente regola:

Regola S9 - Tieni in conto che ci può essere una tensione tra esigenza di trasparenza della Società e esigenza di opacità delle aziende per motivi di concorrenza; le istituzioni nazionali e comunitarie devono affrontare e risolvere queste tensioni

Abbiamo visto nel Capitolo 9 che la Unione Europea ha definito ambiti in cui vieta di produrre e usare modelli basati su Machine learning, e altri ambiti in cui a seconda del livello di rischio definisce regole di progetto da rispettare.

Regola S10 - Quando la impossibilità di garantire la equità e rimuovere tutte le precedenti distorsioni è inattuabile, valuta i rischi e i benefici sociali che possano portare a decidere di evitare l'uso dei modelli predittivi.

Nei progetti in corso in diverse amministrazioni italiane, ad esempio nel Ministero di Giustizia, sta emergendo una linea strategica per cui i modelli classificatori e predittivi non devono essere adottati, ad esempio, nella formazione delle sentenze penali e civili.

Regola S11 – Se possibile, prova a confrontare il risultato prodotto da un modello predittivo con il risultato prodotto da un umano; insomma, sperimenta e confronta, prima di adottarlo.

Metti a confronto tecnologie e persone, e vedi cosa ne viene fuori. Certe volte la tecnologia, in grado di elaborare una quantità immensa di casi, può essere più performante degli umani; oppure gli umani possono adottare intuizioni e soluzioni che non potrebbero mai venire in mente a una tecnica predittiva….

Legata alla precedente è la prossima regola:

Regola S12 – Sperimenta e fai sperimentare soluzioni "ibride" in cui essere umano e modello possano interagire in modo trasparente per migliorare l'equità del risultato complessivo

E’ poi importante che l’equità sia raggiunta nei servizi delle Pubbliche Amministrazioni e dei privati.

Regola S13 - Pretendi come cittadino che lo Stato metta a disposizione sistemi di verifica della equità nei processi amministrativi e imponga alle aziende analoghi sistemi trasparenti di autoverifica nei servizi venduti ai clienti.

Abbiamo commentato la precedente regola nel Capitolo 9 con riferimento alle norme emanate dalla Unione Europea.

Regola S14 - Confronta sempre la trasparenza che puoi ottenere da modelli predittivi basati su tecniche di Machine learning rispetto alle procedure portate avanti da umani, e non dare per scontata la superiorità delle une verso le altre, in entrambe le direzioni.

Le decisioni prese da modelli classificatori e predittivi possono essere viziate da tante forme di iniquità, ma, se noi lo vogliamo, esistono i metodi che permettono di valutarle e mitigarle. *Ragionare su questi metodi ci permette di capire meglio anche le decisioni prese da umani, e scoprire tutti gli elementi di debolezza e di distorsione insite nelle une e nelle altre.*

Regola S15 – Quando progetti un modello classificatorio o predittivo che usi tecniche di apprendimento, le regole di equità non devono essere considerate solo alla fine della realizzazione, ma, piuttosto, come una specifica che orienti il progetto fin dalla fase iniziale. Coinvolgi sempre gli utenti finali nella definizione delle specifiche di progetto.

Questa regola esprime il principio detto *Fairness by design*, equità nel corso del progetto, che suggerisce di procedere nella progettazione del modello coinvolgendo fin dall’inizio gli utenti finali, e conducendo le varie fasi e decisioni progettuali *assumendo la equità come vincolo* nella scelta dei dati di addestramento e nella produzione del modello. Le norme legali devono seguire lo stesso approccio, dovrebbero essere emanate prima della progettazione e produzione di modelli e vanno considerate come un requisito di partenza nel progetto (Responsibility by design).

La prossima regola si occupa del riuso dei modelli in contesti diversi da quelli in cui sono stati inizialmente concepiti.

Regola S16 – Quando applichi a un contesto, per esempio la determinazione del rischio di recidiva, un modello sviluppato per un altro contesto, non applicarlo acriticamente, modifica e adatta il modello al nuovo contesto.

Questa regola si applica quando le aziende che sviluppano modelli predittivi riusano i modelli prodotti in precedenza. Alcuni degli esempi che abbiamo fatto nel libro mostrano che le caratteristiche usate in un caso, e la tecnica di costruzione del modello predittivo, possono non adattarsi a un altro contesto; se ciò viene fatto, c’è un rischio che la accuratezza e la equità cambino completamente. *Pensiamo, ai contesto della recidiva e dei prestiti bancari e alle alle diverse caratteristiche rilevanti per la determinazione del rischio nei due casi.*

Regola S17 – Sensibilizza e forma il cittadino consapevole ad applicare queste regole nella vita di ogni giorno

Ma questa non è una regola, è una metaregola, insomma una regola sulle regole!

Certo! Vedo che hai letto fin qui! Uno dei compiti dello Stato, nell'epoca dei dati digitali, è quello di garantire ai propri cittadini e gruppi sociali la formazione e sensibilizzazione su come i modelli vengono usati nella produzione di decisioni che incidono sulla loro vita, e nella produzione ed erogazione di beni e servizi. E' compito dello Stato cogliere questa esigenza e promuovere un cambiamento culturale. Abbiamo visto che l'equità è definita nella Costituzione italiana, all'articolo 3!

Cittadini più consapevoli possono comprendere meglio e porsi in condizione di pari dignità con i modelli con cui consapevolmente o inconsapevolmente interagiscono. La scelta informata può diventare una forma di garanzia e soglia di attenzione verso la percezione di abusi e il perseguimento di cause e azioni legali collettive.

10.4 Conclusioni – Il Caso ProPublica

Siamo alla fine; cosa pensi, ti è rimasto qualcosa? Prova a esprimerlo pensando al caso della recidiva e dello strumento Compas, ma prova a pensare anche in termini generali…

Devo riordinare le idee…. Ma provo.

Ho scoperto queste nuove tecnologie dei modelli basati su apprendimento, che permettono di classificare e prevedere.

Ho capito che i modelli classificatori e predittivi ci possono essere di aiuto in tante cose della nostra vita, importanti e meno importanti. E ho anche capito che queste tecnologie, che all'inizio appaiono neutrali (cosa c'è di più neutrale di un albero di decisione?) poi però, quando le usi, assorbono, come una spugna, il mondo reale da cui apprendono.

Lo assorbono e lo rappresentano in termini di un secondo mondo, il mondo digitale, anzi, di uno dei tanti mondi digitali che rappresentano lo stesso mondo reale.

L'apprendimento è fatto a partire dai dati del passato. Quindi è per sua natura un apprendimento che non innova, è un apprendimento potremmo dire, conservativo, senza curiosità.

Occorre verificare che il campione sia effettivamente rappresentativo dell'universo e non sia distorto, e, in questo caso, va corretto.

Occorre verificare l'accuratezza del modello classificatorio.

Occorre verificare tutte le regole tecnologiche, ma per fare questo devono intervenire i tecnici, gli scienziati che producono i modelli. I tecnici devono trovare il modo di parlare con i cittadini "normali" devono trovare il linguaggio per farsi comprendere.

Occorre applicare le regole sociali, quelle sono responsabilità di tutti noi.

Riguardo al caso Compas, i nuovi detenuti che fanno domanda di libertà provvisoria sono valutati, come notavo poco fa, sulla base del comportamento passato degli elementi simili.

Se nel passato gli Afro-Americani hanno commesso un maggior numeri di reati, e conseguenti recidive, non è detto che si debba assumere che questo sarà anche per il futuro.

L'arresto non è ancora la condanna, e quindi non corrisponde ad aver commesso il reato, solo ad essere indiziato di aver commesso un reato.

La società deve operare non solo reprimendo i reati, ma nell'ottica di rimuovere le condizioni sociali in cui il reato è maturato.

Il carcere che non dà prospettive, che non rieduca, porta in misura maggiore una volta usciti a commettere nuovamente reati.

Occorre verificare che il feedback non porti ad una spirale di arresti crescenti nelle aree dove risiedono i soggetti classificati ad alto rischio.

E' necessario considerare tutta la vita della persona che fa domanda di libertà provvisoria, il suo percorso formativo, le vicende sociali che ha vissuto al di fuori della esperienza carceraria, le condizioni carcerarie in cui ha scontato la pena.

Ma tutto questo non basta. Mi ha colpito la sezione 9.6 sul cosa e sul perché, sulla differenza tra osservare il mondo e intervenire sul mondo. E' necessario intervenire sulla realtà, non basta osservare.

E' necessario sperimentare nuovi modi di concepire la vita negli istituti penitenziari, nuove forme di attività nel carcere.

E' necessario, ad esempio, mettere a confronto istituti penitenziari con una gestione tradizionale della pena e istituti penitenziari dove di sperimentano forme innovative di attività sociali, artistiche, formative.

Insomma, formulo io questa….

Regola finale – Raggiungere la equità è un processo senza fine. L'equità non potrà mai essere raggiunta, ma, se lo vogliamo, potremo avvicinarla

E' così, per la equità, vale la stessa visione che Umberto Eco esprimeva verso il raggiungimento della *verità*: nella nostra vita difficilmente potremo ricostruire la verità di tutti i fatti che accadono, possiamo però *avvicinare* la verità. Nella nostra vita difficilmente potremo raggiungere l'equità, potremo però, se lo vogliamo, *avvicinare* l'equità, migliorare l'equità, mitigare le iniquità.

Ho capito, grazie…

.

Io, da parte mia, continuo a esplorare……

Appendice 1 - Ambienti didattici e ambienti di misurazione e mitigazione

In questa appendice descrivo due tipi di ambienti utilizzabili per approfondire i temi sviluppati in questo libro:

- gli ambienti didattici per l'apprendimento, dotati di studi di caso che utilizzano visualizzazioni, orientati sia a discenti curiosi, ma non esperti di programmazione, che a utenti che conoscono linguaggi di sviluppo di software, e
- ambienti di algoritmi di misurazione e mitigazione della equità, orientati in prevalenza a docenti e studenti che vogliano sperimentare i concetti forniti in questo libro, e a utenti sviluppatori di software.

Si è cercato perciò di coprire tutte le esigenze e le curiosità di lettori che vogliano approfondire i temi del libro. Tutti i siti citati sono stati verificati nel febbraio 2022.

Riporto direttamente i testi dai siti citati.

1. Google

Google ha sviluppato ambienti didattici per approfondire l'apprendimento del Machine learning e della analisi di equità dei modelli predittivi, dotati di studi di caso, visualizzazioni, esercizi svolti, e ambienti di misurazione e mitigazione della iniquità.

1.1. Ambiente didattico di Google

https://ai.google/education

Vediamo i moduli più vicini a questo corso, con descrizioni dei contenuti basate su quelle del sito.

AI for social good guide - Questa guida aiuta le organizzazioni non profit e le imprese sociali a apprendere come applicare i modelli alle sfide sociali, umanitarie e ambientali.

Che tu sia un principiante o interessato a migliorare le tue abilità, qui ti aiuteremo a comprendere i tipi di problemi che la tua organizzazione può risolvere con il Machine learning, a imparare come identificare e preparare le fonti dei dati di addestramento e a sviluppare e utilizzare il Machine learning in modo responsabile.

The People + AI Guidebook - E' un insieme di metodi e approfondimenti nell'ambito del Machine learning basati sulla esperienza di oltre cento googler, esperti del settore e su ricerche accademiche.

Tic-Tac-Toe, the hard way - Un podcast in cui un docente e un ingegnere del software esplorano le scelte umane che influenzano i sistemi di apprendimento.

AI Explorables. Grandi idee nell'apprendimento automatico, spiegate semplicemente - L'uso in rapido aumento dell'apprendimento automatico solleva domande complesse: come possiamo sapere se i modelli sono equi? Perché i modelli fanno le previsioni che fanno? Quali sono le implicazioni sulla privacy dell'inserimento di enormi quantità di dati nei modelli? Questa serie di saggi interattivi e senza formule ti guiderà attraverso questi importanti concetti.

Fairness - Come abbiamo visto nel libro, la valutazione responsabile di un modello predittivo richiede molto di più del semplice calcolo delle metriche di accuratezza. Prima di mettere in produzione un modello, è fondamentale controllare i dati di addestramento e valutare le previsioni per le distorsioni. Questo modulo esamina diversi tipi di bias umani che possono manifestarsi nei dati di addestramento; quindi, fornisce strategie per identificarli e valutarne e mitigarne gli effetti.

Responsible AI practices - Lo sviluppo dell'Intelligenza artificiale sta creando nuove opportunità per migliorare la vita delle persone in tutto il mondo, dal lavoro all'assistenza sanitaria all'istruzione. Sta inoltre sollevando nuove domande sul modo migliore per integrare equità, interpretabilità, privacy e sicurezza in questi sistemi. Nel sito vengono mostrare alcune applicazioni in questa direzione.

Machine learning glossary - Definisce un ampio insieme di concetti utilizzati nel Machine Learning e sul tema della fairness. Talora i nomi dei concetti sono diversi da quelli adottati in questo libro.

1.2. Ambiente di misurazione e mitigazione della iniquità

I link sono

https://ai.googleblog.com/2020/11/mitigating-unfair-bias-in-ml-models.html

https://www.tensorflow.org/responsible_ai/model_remediation/min_diff/tutorials/min_diff_keras

2. IBM

IBM ha sviluppato *AI Fairness 360,* un ambiente didattico per la sperimentazione di tecniche di Machine learning e di misurazione e mitigazione dei bias nel modelli predittivi.

AI Fairness 360 (AIF360) è un insieme di applicazioni e studi di caso (toolkit) open source completo di metriche sviluppato per verificare la presenza di distorsioni indesiderate nei dati e nei modelli di apprendimento automatico e algoritmi per misurare e mitigare tali distorsioni. E' stato progettato per creare una comunità di utenti per contribuire a creare fiducia nel Machine Learning e rendere il mondo più equo per tutti. Il link al sito è

https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

I progettisti di AI Fairness 360 partono dalla osservazione, ampiamente sviluppata in questo libro, che i modelli di apprendimento automatico sono sempre più utilizzati per prendere decisioni ad alto rischio sulle persone. Sebbene l'apprendimento automatico, per sua stessa natura, sia sempre una forma di bias statistico (chiamata nel libro bias tecnologico) perché la presenza di falsi positivi e falsi negativi è ineliminabile, la discriminazione diventa discutibile quando pone alcuni gruppi privilegiati in vantaggio sistematico e alcuni gruppi non privilegiati in svantaggio sistematico. La distorsione nei dati di addestramento, dovuta ai fenomeni analizzati in particolare nel Capitolo 9, produce modelli che manifestano distorsioni indesiderate.

La versione iniziale (2021) del toolkit AIF360 utilizza Python come linguaggio di sviluppo, e contiene nove diversi algoritmi, sviluppati dalla più ampia comunità di ricerca per mitigare le distorsioni indesiderate, secondo il ciclo di vita mostrato in Figura A1.1.

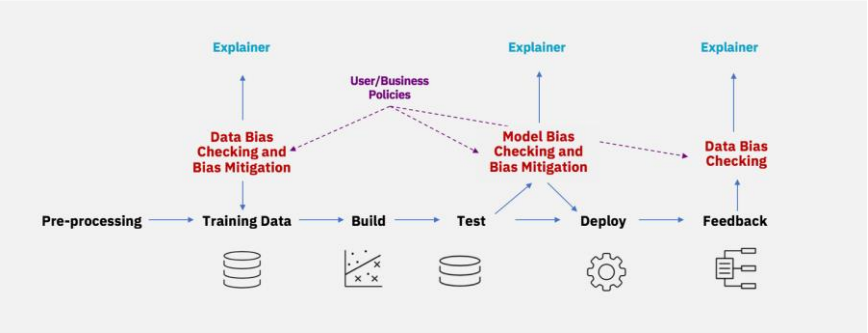


Figura A1.1 – Il ciclo di vita della misurazione e mitigazione delle distorsioni nel toolkit AIF360.

Come si vede, fin dalla fase iniziale di addestramento vengono affrontati i bias causati dai dati (si veda il Capitolo 9), in accordo alle indicazioni (policies) provenienti dagli utenti e, nel caso di utilizzo nel privato, delle politiche aziendali. Fin da questo passo le scelte fatte nel toolkit vengano inviate a un modulo chiamato Explainer, che svolge il compito di spiegare agli sviluppatori e agli utenti le modalità con cui è stata effettuata la verifica dei dati e la mitigazione.

Nel passo successivo vengono considerati sui dati di test i bias derivanti dal modello, anche in questo caso sulla base delle scelte degli utenti. Il ciclo di misurazione e mitigazione prosegue anche nella fase di feedback.

Insomma, il toolkit AIF360 è coerente con l'approccio descritto in questo libro per lo sviluppo, misurazione e mitigazione di modelli predittivi.

Nel sito

http://aif360.mybluemix.net/

sono disponibili programmi e studi di caso sia in Python che in R, i due linguaggi più usati per realizzare modelli basati su apprendimento, per produrre modelli e verificarne la equità secondo diverse definizioni e mitigarne gli effetti.

3. Microsoft

L'ambiente Fairlearn (https://fairlearn.org/) ha due componenti:

- fornisce metriche per misurare il livello di iniquità di gruppo del modello che si sta considerando, per confrontare diverse tecniche (ad esempio alberi di decisione e foreste casuali), per la produzione di modelli in termini di diverse metriche di equità e accuratezza.
- Fornisce algoritmi per mitigare la iniquità, per diversi tipi di iniquità e tecniche.

I gestori del sito Fairlearn consigliano coloro che intendano imparare a usare il sito a consultare prima la guida utente e successivamente il tutorial Quickstart disponibile sul sito.

Inoltre chiariscono che la equità è un concetto socio-tecnico e che non può essere risolto solo con strumenti come Fairlearn. Fairlearn, inoltre, può essere usato sia per modelli predittivi binari (si/no, rischio alto/rischio basso), sia per modelli con un output definito su un dominio continuo (cosiddetti modelli basati sulla regressione).

4. Sito di animazioni visuali in italiano

Il sito

http://www.r2d3.us/una-introduzione-visuale-al-Machine-learning-1/

rappresenta un bellissimo strumento didattico che fornisce tante animazioni con lo scopo di introdurre molte delle tematiche affrontate in questo libro. Lo stile didattico adottato è simile a quello di questo testo: coinvolgere l'utente nella comprensione e formazione dei concetti attraverso esempi intuitivi, così che la loro definizione e le loro proprietà arrivino al discente quando questi ha già sviluppato un quadro cognitivo autonomamente formato. L'uso delle visualizzazioni amplifica significativamente l'efficacia del metodo didattico.

Appendice 2 – Definizioni dei termini principali usati nel libro

Nelle definizioni che seguono adotteremo spesso, per spiegarle meglio, esempi. Ciò non deve distorcere il significato generale del termine, cadendo in quello che è chiamato bias da esempi.

Accuratezza - è, ad esempio nel caso di test di positività per un virus, il rapporto tra le persone positive al test, e che sono malate del virus, più le persone negative al test, e che non sono malate del virus, rispetto al totale delle persone che hanno fatto il test.

Albero di decisione – Insieme di nodi e relative condizioni logiche (ad esempio, è una donna?) che a partire da un nodo origine producono ramificazioni che terminano con nodi foglia. Sono usati per costruire modelli classificatori o predittivi.

Bias, o distorsione o discriminazione, l'opposto di equità – Comportamento di un modello predittivo (ad esempio nell'attribuire un livello di rischio alto) per cui il modello sfavorisce sistematicamente alcuni individui o gruppi di individui rispetto ad altri individui o gruppi di individui.

Campione – Insieme delle persone o altri aspetti del mondo a cui fanno riferimento i dati di ingresso al modello classificatorio/predittivo.

Caratteristica – una proprietà dei dati che descrive la popolazione in ingresso al modello classificatorio/predittivo.

Caratteristica sensibile - Caratteristica delle persone, idonea a rivelare l'origine razziale ed etnica, le convinzioni filosofiche, religiose o di altro genere, le opinioni politiche, l'adesione a sindacati, partiti, associazioni od organizzazioni a carattere filosofico, religioso, politico o sindacale, nonché lo stato di salute e la vita sessuale.

Caratteristica esplicativa (o dati sensibili) – E' la caratteristica, tra le tante, che viene scelta (ad esempio. nel caso di un concorso di ammissione alla Università, il corso di laurea) per comprendere se la discriminazione presente nei dati sia spiegabile senza invalidare la equità, ma unicamente sulla base della informazione obiettiva portata dalla caratteristica.

Correlazione – E' una relazione tra due caratteristiche, tale che a ciascun valore della prima corrisponda un valore della seconda, seguendo una certa regolarità. La correlazione non dipende da un rapporto di causa-effetto, quanto dalla tendenza di una caratteristica a cambiare in funzione di un'altra.

Dati di addestramento – Dati che vengono forniti a una tecnica di apprendimento per costruire un modello classificatorio/predittivo.

Dati di test - Dati su cui si verifica l'accuratezza e la equità di un modello classificatorio/ predittivo.

Discriminazione spiegabile – Discriminazione presente nei dati di addestramento (ad esempio, le diverse percentuali di ammessi alla Università tra uomini e donne) che è giustificabile sulla base, ad esempio, delle diverse percentuali di partecipanti, ma che non viola, nell'esempio, la parità demografica.

Elemento (di un campione o di un universo) – La singola persona o altro aspetto del mondo compresa nell'universo o nel campione.

Entropia – nel contesto delle tecniche di apprendimento, esprime come siano separati gli elementi di una popolazione rispetto ai valori di una caratteristica.

Equità sociale e nei modelli classificatori/predittivi – Caratteristica delle decisioni prese in ambito pubblico e privato, con e senza uso di modelli predittivi, di essere privo di bias o discriminazioni.

Equità comparativa - Fonda il giusto trattamento di una persona, o di un gruppo di persone, sulla base di un confronto con altre persone, o altri gruppi di persone.

Equità non comparativa - Fonda il giusto trattamento di una persona, o di un gruppo di persone, senza considerare come vengono trattate le altre persone o gruppi di persone, ma solo sulla base di un criterio universale di riferimento.

Foresta casuale – Un insieme di alberi di decisione, in cui ciascun albero è costruito a partire da un albero iniziale (o un insieme di alberi iniziali) sulla base di un procedimento casuale nei dati di input considerati e nelle caratteristiche.

Guadagno informativo – E', assumendo come esempio un modello che prevede un rischio, il maggior grado di suddivisione tra elementi a basso rischio e elementi ad alto rischio che l'albero di decisione manifesta nell'insieme delle sue foglie rispetto all'albero considerato in precedenza.

Intelligenza Artificiale - Un qualunque sistema che percepisce l'ambiente in cui opera, ed effettua azioni che massimizzano la possibilità di raggiungere i propri obiettivi. Comprende oltre il machine learning, la logica simbolica, la robotica, l'elaborazione del linguaggio naturale, i modelli conessionistici, le reti neurali, le reti di agenti intelligenti.

Machine learning – Insieme di tecniche con cui si riproduce il ragionamento umano creando procedure che apprendono in base ai dati che utilizzano.

Machine learning da esempi o supervisionato – alla tecnica di apprendimento (ad esempio un albero di decisione) sono forniti, come esperienza, dati di input di un insieme di caratteristiche e

rispettivi dati di output ad essi collegati tramite una funzione che associa a ciascun input il corrispondente output.

Machine learning non supervisionato – alla tecnica di apprendimento sono forniti, come esperienza dati di input senza che questi siano associati a valori di output, ed essa ha il compito di riconoscere schemi/strutture ricorrenti in un insieme di dati.

Machine learning di tipo ensemble – Tecnica di apprendimento che è il risultato dell'uso congiunto di diverse tecniche (ad esempio un albero di decisione e una foresta di alberi casuali) e in cui l'esito della previsione nasce dal prendere in considerazione l'insieme degli esiti delle tecniche, eventualmente in modo pesato.

Modello classificatorio o classificazione – Ha lo scopo di stimare il valore di una caratteristica y (ad esempio l'età) a partire dai valori che assumono un insieme di caratteristiche x1, x2, .., xn. La caratteristica y è anche detta caratteristica obiettivo.

Modello predittivo – Modello classificatorio in cui la caratteristica obiettivo y riguarda un fenomeno ignoto, che non è osservabile nel presente, ovvero che accadrà in futuro.

Parità demografica – La proprietà per cui, ad esempio tra uomini e donne in un concorso deciso da un modello classificatorio, i positivi del modello (nell'esempio, i vincitori del concorso) siano in pari misura, ad esempio, donne e uomini, perché donne e uomini sono presenti in pari misura nella popolazione.

Percentuale di falsi negativi – Una misura di equità in cui si confrontano due gruppi o due elementi sulla base della percentuale di previsioni negative (ad esempio, l'essere ad alto rischio) sbagliate rispetto all'insieme di casi considerato come veri.

Percentuale di falsi positivi - Una misura di equità in cui si confrontano due gruppi o due elementi sulla base della percentuale di previsioni positive (ad esempio, l'essere a basso rischio) sbagliate rispetto a un insieme di casi considerato come veri.

Precisione – E', ad esempio nel caso di test di positività per un virus, il rapporto tra le persone positive al test, e che sono malate del virus, e il totale delle persone che sono risultate positive al test.

Recall o Sensitività – E', ad esempio nel caso di test di positività per un virus, il rapporto tra le persone positive al test, e che sono malate del virus, e il totale delle persone che sono malate del virus, in cui sono incluse le persone che sono malate ma sono risultate negative al test.

Specificità – E', ad esempio nel caso di test di positività per un virus, il rapporto tra le persone negative al test e che non sono malate del virus e il totale delle persone che non sono malate del virus, in cui sono incluse le persone che non sono malate del virus ma sono risultate positive al test.

Spiegabilità – Capacità di esprimere in modo comprensibile il procedimento usato da una tecnica di apprendimento nel produrre un modello predittivo. Può essere locale, quando ci si riferisce a un singolo valore di output (ad esempio il livello di rischio di una specifica persona), globale quando si riferisce all'intero insieme dei valori dell'output.

Tecnica di apprendimento - Tecnica (ad esempio un albero di decisione o una foresta casuale di alberi) che permette di costruire un modello classificatorio o predittivo.

Tecnologia – E' la somma di tecniche, competenze, modelli, metodi, processi e dati digitali usati nella produzione di beni e servizi o nel raggiungimento di obiettivi, come ad esempio la indagine scientifica. La tecnologia può consistere nella conoscenza di tecniche, processi e simili, ovvero può essere integrata in macchine, al fine di permettere il loro utilizzo senza una conoscenza dettagliata del loro funzionamento. I sistemi che applicano la tecnologia trasformando un input in accordo alla funzione svolta dal sistema, e producendo un risultato, sono chiamati *sistemi tecnologici*.

Appendice 3 – Per approfondire

Propongo un insieme di testi per il lettore che voglia approfondire i concetti di questo libro.

Asilomar AI Principles, 2017, vedi https://futureoflife.org/2017/08/11/ai-principles/

Solon Barocas, Moritz Hardt, Arvind Narayanan – Fairness and Machine learning, Limitations and Opportunities, disponibile nel Web, 2021.

Carlo Batini, Federico Cabitza, Paolo Cherubini, Anna Ferrari, Roberto Masiero, Andrea Maurino, Matteo Palmonari, Fabio Stella, – La Scienza dei dati, pubblicato con licenza Creative Commons, liberamente scaricabile dal link https://boa.unimib.it/handle/10281/295980

Jamie Berryhill, Kévin Kok Heang, Rob Clogher, Keegan McBride Hello, World: Artificial Intelligence and its Use in the Public Sector, OECD Working Papers on Public Governance No. 36, 2019.

Carlo Casalone, Luciano Floridi, Laura Palazzani, Renzo Pegoraro, Francesca Rossi, Roberto Villa, Human Centric AI: From Principles to Actionable and Shared Policies, G20 Insights, September 2021.

Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Fostering a European approach to Artificial intelligence Proposal of a regulatory framework in AI and revised coordinated plan on AI, 21 aprile 2021.

Pedro Domingos - L'algoritmo definitivo. La macchina che impara da sola e il futuro del nostro mondo, Bollati Boringhieri, 2016.

Luciano Floridi, Federico Cabitza – Intelligenza Artificiale, Bompiani, 2021.

Stephen Gaukroger – Objectivity, A very short introduction – Oxford University Press, 2012.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi - A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys, 2018.

Deborah Hellman - When Is Discrimination Wrong? – Harvard University Press, 2008.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum - Algorithmic Fairness: Choices, Assumptions, and Definitions - Annual Review of Statistics, 2021.

OECD Framework for the Classification of AI Systems, OECD Digital Economy papers, February 2022 No. 323.

Judea Pearl - The Book of Why: The New Science of Cause and Effect, Penguin, 2019

Frederick Schauer - Profiles, Probabilities, and Stereotypes – Harvard University Press, 2009

Stéphan Vincent-Lancrin and Reyer van der Vlies Trustworthy artificial intelligence (AI) in education: promises and challenges, OECD Education Working Paper No. 218, 2020