Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica
Tesi di Laurea

L'ETICA NELL'INTELLIGENZA ARTIFICIALE: IL PROBLEMA DI MISURARE LA FAIRNESS NEI SISTEMI GOVERNATI DAI DATI

ETHICS IN ARTIFICIAL INTELLIGENCE: THE CHALLENGE OF MEASURING FAIRNESS IN DATA-DRIVEN SYSTEMS

EDOARDO SARRI

Relatrice: *Maria Cecilia Verri*
Correlatore: *Massimiliano Mancini*

Anno Accademico 2023-2024

Edoardo Sarri: *L'etica nell'Intelligenza Artificiale: il problema di misurare la fairness nei sistemi governati dai dati*, Corso di Laurea in Informatica, © Anno Accademico 2023-2024

"Per ogni volta che mi hanno detto:
"ma chi te lo fa fare?!"

RINGRAZIAMENTI

Le tesi sono noiose: nessuno sano di mente dopo averne letta una direbbe "ma che bello, vi prego datemene un'altra che mi sono proprio divertito". Visto questo, voglio prendermi in tutto questo lavoro studiato, corretto, riletto e sistemato, una parte in cui scrivere quello che mi viene, senza che nessuno debba controllarlo.

Si dice che nella tesi ci debbano essere i famosi ringraziamenti, qualche riga in cui si elencano le persone importanti e che ci sono state in questa avventura. Il mio percorso è stato lungo, totalmente fuori di testa e che probabilmente pochi farebbero, e per questo devo molto alle persone presenti nelle prossime righe.

Grazie alla mia Zia, che da quando sono nato si prende cura di me senza considerarmi un nipote, ma un figlio, e che come tale mi tratta. Grazie al mio Zio, che da quando è arrivato mi aiuta e consiglia ogni volta che glielo chiedo e soprattutto quando non lo faccio.

Grazie alla mia Nonna Meri, per ogni volta che si è messa le mani nei capelli dalla paura sapendo che ero a giro o in motorino.

Grazie al mio Nonno Romano, che voleva diventassi un ingegnere ma che non sapeva che più o meno informatica e ingegneria informatica sono la stessa cosa (E poi mai dire mai).

Grazie alla mia Cugina, che nonostante teoricamente non lo sia, in pratica è molto di più: è la mia confidente, la persona da cui posso andare se ho un problema e se voglio fare una bella chiacchierata.

Grazie al mio Cugino, per ogni volta volta che mi ha vomitato tutto il vomitabile quando l'ho preso in collo e per cui prometto di esserci sempre ogni volta che avrà bisogno.

Grazie alla mia Fidanzata, che ogni giorno mi fa sentire l'uomo più amato e desiderato di tutto il mondo, e che mi fa vivere in una nuvolina

incredibile. Mi ha incoraggiato ogni giorno e ogni giorno mi spinge a pensare sempre più in grande. Ogni volta che mi guarda negli occhi capisco di essere veramente felice, e che voglio fare tutto quello che posso per farla restare al mio fianco e ricambiare ciò che mi fa provare.

Grazie al mio Amico, per cui non sarò quello con cui fare le feste ma per cui andrei in capo al mondo se potessi aiutarlo. Grazie ai miei amici: a quelli che mi hanno dato una seconda possibilità nonostante abbia fatto il bastando; a quelli che si interessano a me anche se lontani; a quelli che sono stati un punto di riferimento costante quando ero via da casa.

Si dice che nella tesi ci debbano essere le famose dediche, riferite a qualcuno senza il quale questo percorso non ci sarebbe stato. Io non ne ho una, non ne ho due, e indovinate... non ne ho neanche tre. Ci sono quattro persone senza cui probabilmente non sarei qui o non sarei quello che sono.

Non sarei io senza il mio Babbo e la mia Mamma. Hanno avuto il coraggio di lasciarmi andare via a 17 anni per inseguire il mio sogno. Probabilmente sono stati immensamente male ogni singolo giorno; probabilmente sono stati più volte sicuri che questo momento non sarebbe mai arrivato; sicuramente, però, non mi hanno mai fatto accorgere di nulla. Forse ancora non riesco a capire quanto sia stato difficile il loro lavoro, ma riesco a capire che se un giorno mio figlio o figlia mi dirà "sei proprio simile ai nonni" io sarò l'uomo più soddisfatto del mondo.

Non sarei io senza il mio Nonno Sergio, senza quel santo uomo che ogni giorno mi accompagnava agli allenamenti e stava lì ad aspettarmi come se fosse pagato a peso d'oro. In quel periodo tutto ciò mi sembrava scontato, banale, e non ho mai riflettuto sul fatto che potesse finire. Mi ha insegnato il senso del sacrificio e l'importanza di mettere gli altri davanti a sé stessi, perché vedere felice la propria famiglia è la soddisfazione più grande che c'è. Non ho nessun rimpianto, me lo sono goduto fino all'ultimo momento, ma mi manca tantissimo e questo giorno è anche un po' suo.

Non sarei io senza la mia Nonna Anna, quella fantastica vecchietta che si crede un mezzo rottame ma che in realtà mangia la pappa in testa a

tutto il mondo. Non sarei io perché non avrei trovato nessuno che mi avrebbe attaccato la passione per la matematica e per i computer, che mi avrebbe dato il primo telefono e che si sarebbe messa con me a portare cose dal computer della camerina a quello della cantina usando dei floppy. Probabilmente se sono laureato non è merito suo, ma sicuramente se sulla copertina di questa tesi c'è scritto "Scuola di Scienze Matematiche, Fisiche e Naturali", "Corso di Laurea in Informatica", una gran bella parte del merito va a lei.

INDICE

ELENCO DELLE FIGURE

1

INTRODUZIONE

L'intelligenza artificiale (conosciuta e abbreviata con l'acronimo inglese AI) oggi giorno è sempre più presente all'interno delle nostre vite. L'esempio più lampante e attuale probabilmente è rappresentato da chatGPT. Il chatbot, che risponde a domande su ogni argomento come se fosse un esperto di qualunque cosa, non è però l'unico caso significativo in questo contesto: le piattaforme come Netflix, Amazon o Spotify utilizzano l'AI per raccomandarci, sulla base dei nostri interessi, film, serie, prodotti e canzoni; i veicoli a guida autonoma, come le auto sviluppate da Tesla, sfruttano l'elaborazione di dati acquisti da telecamere e sensori per prendere decisioni di guida e navigare in modo automatico con l'obiettivo di raggiungere una destinazione in modo sicuro e confortevole [60]; gli strumenti di traduzione automatica, come Google Translate, analizzano ed interpretano il contenuto di un testo in lingua straniera, offrendo nel tempo soluzioni sempre più precise.

Tutto questo è possibile grazie alle decisioni che gli agenti intelligenti prendono in maniera automatica, processo basato su quella branca dell'intelligenza artificiale detta machine learning (ML). Come avremo modo di approfondire durante questo lavoro, tale disciplina si basa sull'apprendimento automatico delle informazioni sfruttando vaste quantità di dati provenienti da molteplici canali (social media, applicazioni, dati telefonici, transazioni bancarie) e hanno l'obiettivo di rappresentare le nostre vite e il mondo in cui viviamo.

Poiché la presenza di strumenti basati su AI all'interno della nostra

vita sta prendendo sempre più campo, sorge spontaneo chiedersi se le decisioni che questi agenti prendono siano effettivamente corrette. Senza alcun dubbio, l'accuratezza tecnica rappresenta un aspetto fondamentale, indispensabile affinché una nuova tecnologia si diffonda e possa essere adottata. Tuttavia, in questa tesi, intendiamo concentrare la nostra attenzione su un aspetto tanto cruciale quanto delicato, specialmente considerando che l'AI è arrivata al punto in cui prende decisioni al posto nostro: il concetto di *fairness*.

Un algoritmo di machine learning, per quanto complesso e sofisticato possa essere, resta pur sempre un algoritmo. Dalla definizione di algoritmo, cioè "un insieme finito di istruzioni definite e non ambigue che, eseguito a partire da assegnate condizioni iniziali, produce il risultato corrispondente e termina in tempo finito" [107], si possono evidenziare tre elementi costitutivi: l'input, il processo e l'output. All'interno del così detto "fair machine learning", questi tre aspetti conducono a tre problemi diversi: nel primo caso la preoccupazione principale è che l'input, cioè i dati utilizzati per il training, possano contenere dei pregiudizi (bias); in secondo luogo, c'è la questione definita black-box, cioè la trasparenza del funzionamento dell'algoritmo e del suo non determinismo; infine, l'output potrebbe amplificare i bias contenuti nei dati con cui si è addestrato il sistema [53].

Da quanto appena esposto, risulta normale chiederci se un sistema basato sul machine learning agisca sempre in modo equo, onesto e rispettoso dell'uguaglianza e dell'imparzialità. La risposta a questa domanda non è ovvia e, purtroppo, è ben lontana da essere quella desiderata da tutti. Infatti, non sono molti gli esempi di sistemi intelligenti che hanno brillato per i loro comportamenti etici e in linea con i diritti umani. Dal momento che molti ambiti applicativi dell'AI non sono ancora protetti da leggi contro la discriminazione, ci sono sempre di più attori coinvolti nell'interpretazione di come gli algoritmi codifichino i pregiudizi contenuti nei dati e restituiscano risultati in cui questi sono notevolmente amplificati [99]. In questo contesto, risulta interessante osservare e riflettere sulle parole di Stuart J. Russell, professore di informatica presso la University of California: "Poiché Google, Facebook e altre aziende stanno

attivamente cercando di creare una macchina intelligente, una delle cose che non dobbiamo fare è andare avanti a tutto vapore senza pensare ai rischi potenziali. Se si vuole una intelligenza illimitata, è meglio capire come allineare i computer con i valori e i bisogni umani" [15].

Per affrontare e cercare di risolvere i problemi appena discussi, in letteratura è presente un'ampia gamma di proposte riguardanti algoritmi di ML che tengono conto dell'equità. Queste tecniche prevedono sia interventi di pre-processing che di post-processing, rispettivamente sui dati di input e output, oltre a interventi diretti sugli algoritmi di apprendimento automatico stessi [66].

L'obiettivo di questa tesi coinvolge un aspetto fondamentale, che precede tutto ciò che è stato introdotto fin'ora. Se quello a cui miriamo è sviluppare sistemi che rispettino il principio di fairness, dobbiamo innanzitutto trovare un modo per definire questa proprietà. Mentre a livello umanistico e qualitativo lo stato dell'arte non ci pone davanti a molti dubbi, descrivere la fairness in modo matematico e quantitativo, così da poterla misurare all'interno di sistemi basati sul machine learning, risulta un compito assai più arduo [70].

La meta che vogliamo raggiungere alla conclusione di questa tesi, è essere in grado di misurare e stabilire se un dataset soddisfa il criterio fairness, secondo alcune delle definizioni che presenteremo. Vogliamo poter arrivare a questo risultato indipendentemente dall'utilizzo di un modello di machine learning. Scopriremo che che un particolare gruppo di definizioni non coinvolge le previsioni dell'algoritmo, ma si basa soltanto sul dataset considerato. Visto l'elevato numero di metriche proposte in letteratura, ognuna delle quali presenta sfumature specifiche, questo compito è tutt'altro che banale; noteremo che trovare una definizione univoca al concetto di fairness è una sfida tanto significativa quanto cruciale per il progresso nell'ambito del machine learning e dell'intelligenza artificiale. Spesso infatti, risulta più semplice rilevare la presenza di discriminazione piuttosto che la sua assenza. A prova di questo presenteremo un particolare dataset, Adult, che analizzeremo successivamente utilizzando due definizioni. Questo ci permetterà di trarre le nostre con-

clusioni e dimostrare che, anche per casi molto conosciuti e ampiamente utilizzati, è essenziale soffermarsi a riflettere sul fatto che le intelligenze artificiali, che tanto acclamiamo e utilizziamo, non sempre si comportano esattamente come dovrebbero.

1.1 Struttura della tesi

La struttura del lavoro di tesi è la seguente:

- Nel Capitolo 2 forniremo una panoramica degli elementi necessari per il proseguimento del lavoro. Introdurremo i concetti base dell'intelligenza artificiale e del machine learning, esaminando la loro evoluzione storica e esplorando le loro applicazioni. Approfondiremo il funzionamento del machine learning cercando di capire come funziona l'apprendimento automatico, evidenziandone le differenze e portando degli esempi significativi. Successivamente, introdurremo la fairness, concetto chiave di questa tesi. Attraverso la presentazione di due esempi significativi, il sistema americano COMPAS e quello impiegato dalle aziende durante il processo di assunzione del personale, cercheremo di capire quali siano i principali problemi in situazioni simili a queste. Infine porremo l'attenzione sull'importanza della correttezza e sull'assenza di bias nei dati di training, utilizzati per addestrare gli algoritmi di machine learning.
- Nel Capitolo 3, dopo una breve introduzione sulle basi statistiche che ci serviranno, l'obiettivo sarà esaminare, dal punto di vista matematico e statistico, le varie definizioni di fairness presenti in letteratura.
- Nel Capitolo 4 sarà presentato il dataset che utilizzeremo nel seguito, Adult. Dopo un accenno sulla centralità delle variabili sensibili nell'ambito della fairness, esamineremo in dettaglio le caratteristiche del dataset e dei suoi attributi, identificando quelli critici e su cui concentrare l'attenzione.

- Nel Capitolo 5, prima di trarre le nostre conclusioni, valuteremo Adult cercando di capire se soddisfa la proprietà di fairness secondo una specifica classe di definizioni. L'obiettivo sarà quello di fornire una dimostrazione pratica di quanto discusso durante l'intera tesi, cercando di comprendere se gli aspetti etici siano rispettati quanto dovrebbero.
- Nel Capitolo 6 cercheremo, infine, di analizzare quanto ottenuto in precedenza in relazione all'obiettivo che ci eravamo proposti.

2

CONTESTO

2.1 Intelligenza artificiale

L'intelligenza artificiale fonda le sue radici nel campo dell'informatica, ma oggi è diventata una disciplina a sé stante, influenzata da diverse aree del sapere, quali matematica, psicologia e scienza.

Definire cosa sia l'intelligenza artificiale potrebbe essere un compito non banale: se sul termine "artificiale" si può essere in grado di trovare una linea comune e non ambigua, il concetto di "intelligenza" può sollevare interrogativi ben più complessi. Nel corso della storia molti studiosi hanno affermato che l'intelligenza artificiale non si limita a studiare solo gli aspetti biologici dell'intelligenza umana, ma vuole cercare di osservare, comprendere e replicare il processo decisionale, e quindi la parte computazionale, del cervello umano [72]. Con questa idea l'intelligenza artificiale può essere vista come un insieme di metodi e tecniche il cui scopo è quello di creare sistemi capaci di elaborare informazioni e risolvere problemi in maniera simile a come farebbe la mente umana e che siano in grado successivamente di collaborare o addirittura competere con essa [50].

Entrando maggiormente nel dettaglio, una delle definizioni storiche che ha segnato un punto di svolta nella storia dell'intelligenza artificiale è quella espressa da Alan Turing. Passato alla storia con la frase "Can machines think?" [106], ha proposto il primo e forse più famoso test per la misurazione dell'intelligenza di una macchina: il test di Turing. Secon-

do questo test, una macchina può considerarsi intelligente se, durante un'interazione testuale con un giudice umano, quest'ultimo non riesce a distinguerla da un essere umano [50, 72].

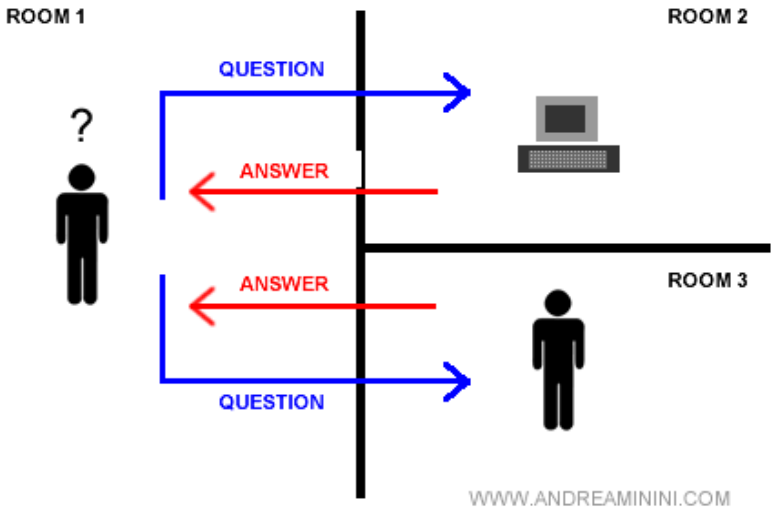

Figura 1: Funzionamento del test di Turing [7]

Oltre al test di Turing, un'altra definizione fondamentale è quella basata sul concetto di agente intelligente: un "agente" è un software che percepisce l'ambiente circostante tramite sensori e agisce su di esso tramite attuatori; il termine "intelligente" si riferisce alla capacità di selezionare un'azione che rende massima una misura di performance [48].

Una definizione più recente che indirizza l'attenzione sulla complessità dell'intelligenza artificiale è quella proposta dall'AI HLEG ("Artificial Intelligence High-Level Expert Group") durante la conferenza europea del 2018. Il gruppo di esperti nominato dalla commissione europea per "fornire consulenza sulla strategia di intelligenza artificiale" [24], ha affermato che l'intelligenza artificiale è un sistema composto da hardware e software che, sfruttando modelli matematici, interpreta ed elabora i dati provenienti dall'ambiente circostante al fine di svolgere compiti e raggiungere obiettivi prefissati [20]. Questa definizione sottolinea l'importanza dei modelli matematici per l'elaborazione di dati, mettendo quindi

in risalto l'ampio spettro di metodologie e approcci diversi utilizzabili nello sviluppo.

Una volta definita l'intelligenza artificiale, sono molte le domande che ci possiamo porre. Una di queste riguarda il dominio dell'uomo sulla terra: fino ad oggi l'uomo ha sempre avuto il predominio rispetto alle altre specie presenti sul pianeta Terra, ma se l'intelligenza artificiale può pensare e ragionare tanto bene da competere con l'uomo siamo sicuri che questa situazione rimarrà invariata? Sulla base di questa riflessione possiamo definire due tipi di intelligenza artificiale: debole e forte.

L'intelligenza artificiale debole, comunemente nota a livello internazionale come "weak artificial intelligence", si riferisce allo sviluppo e all'utilizzo di agenti intelligenti in settori ben specifici e molto complessi; questi sistemi possono simulare, nel senso di imitare, gli aspetti cognitivi umani, ma non sono in grado di replicarli completamente (ad esempio, non sono capaci di pensare) [63]. Esempi sono i robot, le auto a guida autonoma e gli algoritmi di ricerca intelligenti. L'altro lato della medaglia è l'intelligenza artificiale forte, nota come "strong artificial intelligence", che si riferisce al tipo di intelligenza in grado di superare l'essere umano in ogni aspetto. Ci sono più sfere della mente che possono appartenere all'AI forte: la sfera emozionale, quella sociale, quella sensomotoria e quella cognitiva; sebbene al momento la sfera emozionale non sia contemplata come propria di un robot, si suppone che in un futuro un sistema intelligente possa acquisire la capacità di provare emozioni e sviluppare una coscienza autonoma [46].

Tra le due categorie l'intelligenza artificiale debole è sicuramente quella più sviluppata e, affermazione che può essere formulata forse con po' di sollievo, sembra anche essere quella più facilmente sviluppabile. In questo campo ricerca si sono ottenuti risultati che solo qualche anno fa sembravano molto lontani. Questi risultati oggi possono essere impiegati in settori come finanza, medicina e marketing, solo per citarne alcuni [63, 46].

2.1.1 Storia

L'elevato numero di applicazioni dell'intelligenza artificiale, dalle più banali a quelle più creative e innovative, è oggi sotto gli occhi di tutti. Tuttavia, non possiamo trascurare il lungo percorso e il lento sviluppo, caratterizzato da approcci molto diversi tra loro, che questa disciplina ha attraversato nel corso del tempo.

Uno dei contributi più rilevanti, e che ha posto le basi per i lavori degli anni a seguire, è stato quello di Warren McCulloch e Walter Pitt sulle reti neurali [73], "modelli matematici del sistema nervoso e del comportamento simulati tramite computer" [9]. Questo lavoro ha infatti portato lo scienziato e matematico Marvin Lee Minsky nel 1951 a costruire il primo computer a reti neurali, SNARC [93], capace di replicare una rete di 40 neuroni.

Nonostante il concetto di intelligenza artificiale sia molto discusso e trattato, viene confuso con quello di machine learning; spesso i due termini vengono infatti erroneamente usati in modo interscambiabile. Per fare un po' di chiarezza è importante comprendere che, senza entrare ancora nel dettaglio, mentre l'intelligenza artificiale rappresenta la disciplina di base, il machine learning è una tecnica, o per meglio dire un modello, che ne consente l'applicazione pratica [41]. Come è facile immaginare, anche se il machine learning oggi occupa una posizione dominante in questo settore, non è stato il primo approccio.

Dopo gli anni '60 l'attenzione era quasi totalmente focalizzata verso i così detti sistemi esperti. Nel 1965 il progetto DENDRAL, progettato per mappare la struttura delle molecole al fine di assistere i chimici nell'identificazione delle molecole organiche sconosciute, rappresentò un passo significativo in questo campo e fu considerato il padre di tutti i sistemi esperti [74]. Gli anni '80 furono caratterizzati da un grande entusiasmo riguardo alla possibilità che i sistemi esperti superassero le prestazioni intellettuali umane [39], come dimostrato da una vittoria di una macchina contro un essere umano nel gioco degli scacchi [97].

L'obiettivo dei sistemi esperti era quello di simulare il giudizio e il

comportamento di un essere umano dotato di grande conoscenza ed esperienza in un campo specifico; lo scopo era accedere tramite delle interviste a tali informazioni in modo da permettere di risolvere un determinato problema complesso anche ad un utente meno esperto [105]. Questi sistemi sono basati sulla codifica di regole, procedure e metodi precisi e dettagliati; l'obiettivo non era quello di costruire un sistema che imparasse ed apprendesse dall'ambiente e dall'esperienza nel tempo, ma che riconoscesse situazioni già definite in modo da risolvere istanze di problemi già noti in modo efficiente e deterministico [42].

Negli anni successivi, l'AI trovò applicazione anche nel campo della traduzione automatica, ma tuttavia in questo contesto i risultati furono alquanto deludenti. Un esempio emblematico, ma sicuramente non l'unico che potremmo presentare, fu quello che coinvolse il DARPA, un'agenzia governativa del Dipartimento della Difesa degli Stati Uniti incaricata di sviluppare nuove tecnologie da sperimentare in ambito militare. Il compito assegnato era quello di progettare un sistema in grado di tradurre articoli scientifici dal russo all'inglese. A causa di numerosi fallimenti, che portarono a risultati errati e grotteschi, nel 1966 il governo americano decise di interrompere i finanziamenti per la ricerca.

Oltre a questa situazione e ad altre molto simili, in questo periodo furono anche sollevate le prime critiche pubbliche nei confronti del settore dell'intelligenza artificiale. Hubert L. Dreyfos, prima nel suo articolo "Alchemy and AI" del 1965 e successivamente nel libro "What computer can't do" del 1972, attaccò i ricercatori asserendo che l'AI non fosse realizzabile. Il filosofo statunitense pose l'accento sull'incapacità di realizzare i successi tanto promessi, mettendo quindi in discussione le aspettative e il futuro della tecnologia proposta [47].

Dopo una serie di fallimenti, negli anni '80 si tornò a respirare la stessa aria di innovazione e fiducia che caratterizzò il periodo dei sistemi esperti. Il lavoro di David Rumelhart e James McClelland sulle già citate reti neurali, fece comprendere ai ricercatori che la strada intrapresa poteva essere quella corretta [47]. I due psicologi americani, nel loro tentativo

di spiegare il cervello umano con le reti neurali artificiali, arrivarono a proporre un nuovo modello: il Parallel Distributed Processing (PDP). Questo nuovo approccio permise di costruire e programmare un sistema basato sulla notevole semplificazione della mente umana, le cui caratteristiche chiave di si basavano su due aspetti: le informazioni non dovevano passare obbligatoriamente per la CPU in modo seriale, rallentando inevitabilmente il sistema; le informazioni non erano collocate in un'unica area di memoria, ma bensì erano distribuite su tutti i nodi della rete [92].

Lo stato dell'arte è, al giorno d'oggi, ben più sviluppato e si basa su complessi modelli matematici accuratamente strutturati. Se un tempo veniva richiesto di inserire manualmente tutti i dati, ora i sistemi basati sulla conoscenza attingono le informazioni da grandi database e si affidano agli algoritmi di apprendimento più avanzati. Attualmente, nello sviluppo dell'AI la tendenza è quella di basarsi sui concetti già esistenti, affermati e consolidati: questi sono infatti sfruttati per validare nuove ipotesi oppure per giungere a nuove conseguenze e evidenze sperimentali [47].

2.1.2 Applicazioni

Negli ultimi anni l'intelligenza artificiale ha senza dubbio assunto una posizione di grande rilievo all'interno del mondo scientifico, diventando uno dei settori di ricerca più attivi e in continua evoluzione. In questo contesto possiamo individuare due fazioni ben distinte: da una parte ci sono i suoi sostenitori, che cercano di sfruttarne a pieno il potenziale per ottenere dei benefici; dall'altra parte ci sono i suoi detrattori che ne criticano l'utilizzo, esprimono scetticismo e manifestano timore per un futuro in cui questa nuova tecnologia sembra essere sempre più centrale. È chiaro che le opinioni possono variare in base a molti fattori, come le esperienze personali, le prospettive future individuali e le capacità di non restare inermi di fronte a questo cambiamento. Schierarsi in una delle due fazioni è una scelta sicuramente molto complessa, tanto da risultare quasi impossibile. È mia opinione che molti siano divisi tra la possibilità

di sfruttare le opportunità che una tecnologia così rivoluzionaria mette a disposizione e la preoccupazione di un elevato numero di aspetti negativi, come la scomparsa di determinati posti di lavoro e la violazione della privacy. In un contesto così delicato risulta molto complesso trovare un equilibrio stabile tra i benefici dell'impiego dell'intelligenza artificiale nella vita di tutti i giorni e la tutela dei valori e dei diritti umani.

Figura 2: EU Artificial Intelligence Act [2]

Dal punto di vista politico, il 2017 segna senza dubbio il punto di svolta nello sviluppo e adozione dell'intelligenza artificiale. Putin, il Presidente russo, durante il dibattito internazionale sullo sviluppo dell'intelligenza artificiale ha sollevato l'attenzione dichiarando che "chiunque diventerà il leader in questo ambito diventerà il governante del mondo" [102]; questa affermazione mette in risalto il riconoscimento delle potenzialità attribuite all'intelligenza artificiale. Nello stesso periodo altre nazioni, come l'India e la Cina, hanno dimostrato un crescente interesse nell'argomento: la prima ha infatti istituito una Task Force per studiare gli effetti politici ed economici di questa tecnologia; la seconda ha fissato l'ambizioso obiettivo di diventare una superpotenza in questo campo entro il 2030, investendo in infrastrutture, ricerca e sviluppo [35, 59].

Più recente è invece l'approvazione da parte del Parlamento Europeo dell'"EU Artificial Intelligence Act". Si tratta di una legge che "garantisce la sicurezza e il rispetto dei diritti fondamentali e promuove l'innovazione" [29]. L'obiettivo principale di questa normativa è proteggere i diritti, la democrazia e la sostenibilità ambientale dall'intelligenza artificiale, promuovendone contemporaneamente l'innovazione per assicurare all'Europa un ruolo di rilievo all'interno settore. Il nuovo regolamento proibisce quelle applicazioni dell'AI che potrebbero minacciare i diritti dell'Uomo: la cattura di immagini facciali da internet o da sistemi di

registrazione a circuito chiuso, al fine di creare database di riconoscimento facciale; sistemi che riconoscono le emozioni all'interno di contesti scolastici e lavorativi; sistemi progettati per manipolare il comportamento umano o che sfruttano le vulnerabilità delle persone. Significativi cambiamenti sono stati introdotti anche nell'ambito della trasparenza: ogni contenuto che include immagini generate dall'AI, dovrà essere chiaramente etichettato come tale [29, 27].

L'adozione dell'intelligenza artificiale sta sicuramente crescendo molto rapidamente praticamente in ogni settore. Questo porta le aziende ad introdurre al loro interno strumenti e tecniche innovative, che portano a risultati tangibili da un punto di vista economico.

Un esempio è lo sviluppo dei chatbot, impiegati dalle aziende per assistere la clientela. Fornendo un supporto on-demand, sempre disponibile e immediato per la risoluzione dei problemi più semplici, si riesce a raggiungere l'obiettivo di migliorare l'esperienza dell'utente e allo stesso tempo si aumentano i profitti dell'azienda. Un altro impiego con cui molte persone nell'ultimo decennio sono entrate in contatto è la guida autonoma. I processi di apprendimento automatico e di percezione dell'ambiente circostante tramite sensori, porta i veicoli a prendere decisioni in tempo reale, migliorando la sicurezza sulle strade e offrendo nuove opportunità per le aziende che operano nel settore dei trasporti.

Infine, possiamo citare la così detta Industria 4.0, una rivoluzione che integrando l'Internet of Things (IoT), il cloud computing e l'analisi dei dati, mira a migliorare il processo produttivo. Grazie all'intelligenza artificiale le aziende possono massimizzare i ricavi, abbattere i tempi di produzione e prevedere la domanda del mercato con l'obiettivo, ancora una volta, di aumentare i ricavi.

Sebbene possano essere meno evidenti per una persona che non è direttamente interessata, le applicazioni dell'intelligenza artificiale in campo medico sono di grande rilievo. Le organizzazioni sanitarie hanno una vasta quantità di dati dei pazienti e, in un crescente numero di ambiti, gli algoritmi di machine learning si sono dimostrati più abili

degli umani nell'interpretare tali dati e individuare le relazioni presenti al loro interno [84]. Il loro addestramento su dati molto accurati come sono quelli medici, può portare non solo a una migliore diagnosi e alla cura della malattia, ma anche ad una prevenzione più efficace della salute dei pazienti. Infatti, se un medico può fallire nell'individuare il legame tra una malattia e un particolare risultato di un esame, questo è meno probabile quando l'osservazione viene eseguita da un algoritmo di machine learning accuratamente addestrato per questo scopo [43].

In altre parole, l'intelligenza artificiale offre un grande potenziale alla pratica medica, che se sfruttato potrebbe permettere non solo di migliorare la qualità delle cure mediche, ma anche di ridurre i costi sanitari.

Se consideriamo il concetto di intelligenza artificiale strettamente legato all'intelligenza umana, soprattutto dal punto di vista del funzionamento del cervello, il progetto forse più sensazionale e allo stesso tempo presuntuoso è quello portato avanti da Elon Musk attraverso una delle sue aziende, Neuralink.

L'obiettivo che vuole essere raggiunto entro il 2026 è, infatti, quello di "sviluppare interfacce neurali impiantabili che facciano comunicare il cervello con il computer e aiutino a curare gravi disabilità" [31].

La commissione europea tra gennaio e marzo 2020 ha condotto un'indagine [62] su quasi 10.000 aziende arrivando a conclusioni senza dubbio interessanti. Il 78% delle aziende sotto esame non solo era a conoscenza dell'intelligenza artificiale, ma riconosceva anche il suo potenziale all'interno dei propri processi aziendali; inoltre, circa il 42% di esse in quel periodo aveva già adottato almeno una tecnologia basata sull'intelligenza artificiale, mentre il tasso di adozione di due o più di queste tecnologie scendeva al 25%.

Se da un lato l'intelligenza artificiale si sta sviluppando tanto da sembrare al centro di una quarta rivoluzione industriale [104], dall'altro emergono una serie di problemi che questa tecnologia porta con sé. In particolar modo, dal punto di vista delle aziende, lo studio sopra cita-

to mette in evidenzia una grande difficoltà nell'individuare personale specializzato e competente in un campo sostanzialmente nuovo; inoltre, nonostante sempre più aziende desiderino introdurre strumenti basati sull'intelligenza artificiale, i costi della loro adozione sono ancora troppo elevati rispetto ai benefici che ci si aspetta ottenere [42].

2.2 Machine learning

Tornando al concetto dei sistemi esperti descritti nel paragrafo precedente, nonostante essi siano una soluzione ottima in contesti molto specifici grazie alla loro capacità di essere estremamente settoriali, in molte circostanze non sono effettivamente utilizzabili. L'ostacolo predominante è rappresentato dal loro determinismo: un sistema esperto si basa su regole e relazioni predefinite, che devono essere codificate per intero da un programmatore [65]. Se il numero di queste regole è limitato allora la loro realizzazione è possibile. Ci sono molte situazioni però, soprattutto nei contesti più generici, in cui trovare pattern all'interno di un numero di relazioni e regole molto elevato è praticamente impossibile. In questi casi, se non è possibile codificare algoritmi ad-hoc dove le istruzioni sono strettamente statiche, la soluzione deve essere cercata in altri modelli.

Sebbene descrivere correttamente l'intelligenza artificiale e i suoi obiettivi, come abbiamo già visto, possa essere un compito molto complesso, possiamo sicuramente affermare che un suo aspetto fondamentale, considerato che l'idea generale è quella di assomigliare ad una mente umana e di replicare il suo ragionamento, è l'apprendimento. In questo contesto l'apprendimento gioca un ruolo talmente importante che nel corso del tempo ha dato origine ad una disciplina a sé stante che prende il nome di machine learning [65].

Arthur Samuel fu il primo ad utilizzare il termine machine learning, riferendosi alla branca dell'informatica che sfrutta algoritmi che dettano le regole per la lettura e interpretazione delle informazioni [95].

Per dare un'idea di machine learning senza entrare nel dettaglio possiamo dire che è il rilevamento automatizzato di modelli e pattern significativi presenti nei dati [98]. L'obiettivo principale è quello di modellare una relazione tra dati in ingresso (input) e informazioni ad essi collegate (output); questi modelli poi consentono di prevedere i valori delle variabili desiderate misurando gli elementi osservabili [81]. In breve, il machine learning può essere visto come una branca dell'intelligenza artificiale che si concentra sulla definizione di modelli e sull'implementazione di algoritmi non deterministici che permettono di apprendere dai dati passati, riconoscere pattern e prendere decisioni senza essere stati esplicitamente programmati per farlo.

2.2.1 Storia

La storia del machine learning, come si può immaginare, è strettamente legata a quella dell'intelligenza artificiale, essendo il primo una delle principali tecniche con cui la seconda può essere realizzata.

Un primo approccio all'idea di apprendimento automatico fu seguito da Arthur Lee Samuel, che, come abbiamo già visto, oltre ad usare per la prima volta il termine machine learning con il significato che indicativamente gli attribuiamo anche oggi, nel 1952 inventò il primo programma in grado di apprendere dall'esperienza. La tecnica ideata da Samuel per addestrare il sistema, che oggi sarebbe stata classificata sotto l'apprendimento supervisionato, permise di abbandonare la concezione tradizionale che si aveva di computer, ovvero quella di esecutore di compiti ben schedulati e definiti. Il programma ideato riusciva a migliorarsi progressivamente nel gioco della dama, tramite l'esperienza ottenuta in partite contro se stesso o contro altri avversari [95].

Successivamente, nel 1957, Frank Rosenblatt creò il primo modello basato sulle reti neurali. Il programma prese il nome di Perceptron e sostanzialmente era un classificatore binario, in grado di riconoscere e classificare in due classi nuove informazioni sulla base di un confronto con quanto di più simile avesse in memoria [69].

Un ulteriore passo importante è stato compiuto con lo sviluppo dei primi sistemi di sintesi vocale, in grado di tradurre un testo scritto in inglese in formato audio [108]. L'algoritmo in questo caso fu allenato sulla base di discorsi formali, e il sistema era basato su reti neurali.

Dopo gli anni 2000 per le più grandi aziende mondiali divenne chiaro che l'utilizzo degli algoritmi di machine learning fosse essenziale per gestire e analizzare i big data. Il termine "big data" si riferisce a un enorme insieme di dati caratterizzato non solo dalla sua grandezza, ma anche dalla sua complessità di archiviazione, analisi e visualizzazione [94].

In un contesto di questo tipo, trovare pattern e mettere in relazione informazioni analizzate manualmente senza l'ausilio di strumenti che svolgano il compito in maniera automatica, può risultare molto complesso ed estremamente inefficiente [109]. Sono state proprio le situazioni di questo tipo che hanno permesso nel tempo un matrimonio così solido tra machine learning e big data: da un lato, infatti, il machine learning ha permesso di rendere più comprensibile e trattabile la grande mole di dati che sono raccolti da una grande varietà di dispositivi collocati in tutto il mondo; dall'altra, i big data offrono le informazioni necessarie per addestrare gli algoritmi di machine learning. Questo permette ai sistemi di migliorare costantemente le proprie prestazioni e contemporaneamente permette di comprendere e analizzare sempre meglio i dati e le informazioni a nostra disposizione [68, 54].

2.2.2 Apprendimento

Fino ad ora abbiamo sottolineato più volte la centralità dell'addestramento nel contesto del machine learning. In questo paragrafo cerchiamo di fissare meglio i concetti del suo funzionamento; non scenderemo nel dettaglio di ogni caratteristica perché questo non è l'obiettivo di questa tesi, ma comunque cercheremo di esaminare i concetti principali.

Iniziamo col dire che si può parlare di algoritmo di apprendimento ogni volta in cui siamo di fronte ad un programma che impara da espe-

rienze passate. Esistono due tipi generali di apprendimento: induttivo o deduttivo. Nel primo caso si fa riferimento alla scoperta di regole o leggi generali partendo da esempi specifici; nel secondo caso si parte da un insieme di regole o fatti conosciuti per derivare ipotesi che si adattino al contesto. Quando si parla di machine learning, solitamente si fa riferimento a tecniche di apprendimento induttive [55]; questo vuol dire che una macchina è in grado di generalizzare dalla propria esperienza, cioè può portare a termine in maniera accurata compiti nuovi e mai visti basandosi su informazioni che sono state acquisite precedentemente [81, 49].

Una delle definizioni più citate e più formale di machine learning ci viene data dal professor Tom Mitchel della Carnegie Mellon University: "un algoritmo di apprendimento si dice che sia in grado di imparare da un'esperienza E, con riferimento ad alcune classi di compiti T e ad alcune misure di performance P, se le sue performance nei compiti T, misurate tramite P, migliorano con l'esperienza" [77]. La performance citata da Mitchel è solitamente valutata in modo statistico come la percentuale di campioni correttamente predetti. Tuttavia, c'è un altro modo, in parte opposto al precedente, di calcolare la performance, ovvero come la percentuale di campioni erroneamente predetti. Un aspetto fondamentale di questa definizione è legato al miglioramento continuo, dimostrato solitamente con una misura quantitativa, che l'algoritmo deve avere all'aumentare dell'esperienza.

Entrando più nel dettaglio possiamo dire che l'apprendimento, nel contesto del machine learning, è il processo automatico di creazione del modello, usato poi per compiere predizioni o prendere decisioni; i dati usati durante questo processo di costruzione sono invece comunemente detti dati di addestramento [37].

I metodi di addestramento nel machine learning sono molti e si dividono in più classi a seconda del loro funzionamento [78]. La categorizzazione che indicheremo nel seguito può essere fatta risalire al già citato Arthur Lee Samuel: egli ha distinto i sistemi intelligenti tra quelli che fanno riferimento a dati disponibili da osservare, da quelli che invece

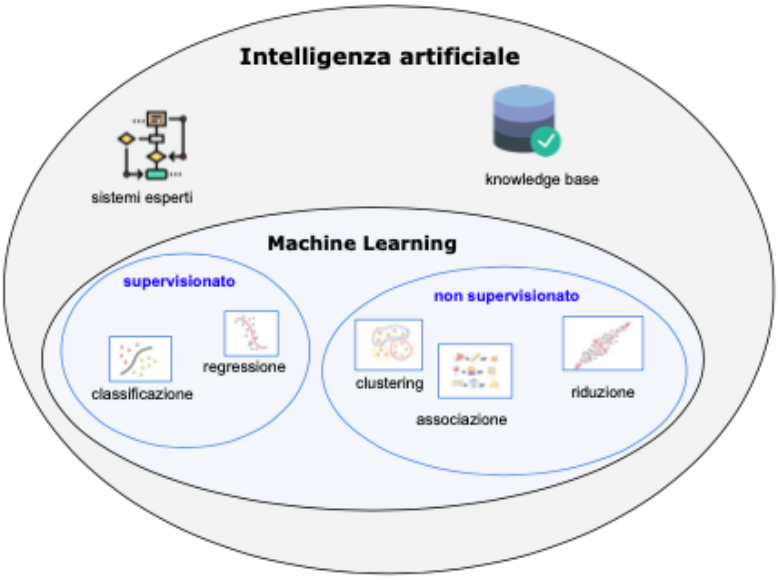

Figura 3: Intelligenza artificiale e machine learning [70]

lavorano in autonomia, cioè senza basarsi su schemi già presenti. L'esperienza a cui sono sottoposti gli algoritmi di addestramento durante il processo di apprendimento è solitamente declinata nel modo seguente:

- Apprendimento supervisionato

  Nell'apprendimento supervisionato i dati di apprendimento sono specificatamente contrassegnati da un insieme finito di numeri interi, noti come etichette o numeri di classe [83]; l'algoritmo, utilizzando i dati di partenza come esempio, dovrà essere in grado di categorizzare nel modo corretto dati mai osservati prima attraverso le etichette di classe.

  In base al tipo di risultato desiderato, possiamo far riferimento ad un'ulteriormente classificazione: gli algoritmi di classificazione e algoritmi di regressione [41]. Negli algoritmi di classificazione l'output consiste in valori discreti. Si parla di classificazione binaria se l'output può avere solo due valori, cioè se ci interessa dividere i dati in due sottogruppi; si parla invece di classificazione multiclasse se sono presenti più valori possibili. L'obiettivo principale di

tali algoritmi è creare un modello che permetta di associare ad ogni input mai osservato una specifica classe [82].

Negli algoritmi di regressione l'output assume valori continui. In questa situazione sono disponibili più variabili indipendenti, dette variabili predittive, e una variabile dipendente, detta variabile target; l'obiettivo del modello è quello di stabilire una relazione tra l'insieme delle variabili indipendenti e quella dipendente, in modo da prevedere il suo valore successivamente [82, 85].

Un esempio di algoritmo di classificazione binaria è il filtro anti-spam presente nelle caselle mail: questo, infatti, suddivide tutte le mail in entrata (campioni ancora mai visti) in due classi, le mail spam e quelle non spam. Se parliamo invece di algoritmi di regressione un esempio lo si può identificare nella predizione del valore di un'azione di un'azienda quotata in borsa: in questo caso stiamo cercando un risultato preciso, che non è identificativo di una classe ma è puramente un'informazione.

- Apprendimento non supervisionato

  Nell'apprendimento non supervisionato i dati di apprendimento non sono stati precedentemente etichettati o annotati da essere umani o algoritmi. In questa situazione non c'è quindi un output previsto, cioè l'algoritmo di apprendimento non ha nessuna indicazione sulla relazione tra informazioni e risultato [54, 100].

  L'obiettivo che si vuole raggiungere è quello di consentire all'agente intelligente di scoprire relazioni nascoste che potrebbero non essere note agli esseri umani o che questi non sarebbero in grado di trovare. In altre parole, l'algoritmo cerca pattern e strutture nei dati senza nessuna guida o informazione predefinita; una volta che viene consegnato un risultato è compito di un supervisore correggerlo o accettarlo [64].

  Un esempio di algoritmo di classificazione non supervisionato è Google News [41]: il motore di ricerca del colosso americano analizza e organizza in modo automatico una vasta quantità di

notizie, provenienti da diverse fonti e trattanti argomenti diversi, con l'obiettivo di capire quali siano rilevanti e quali no. Un ulteriore e interessante esempio è la rimozione del rumore in esami medici [101], come l'ecografia.

- Apprendimento con rinforzo

  L'apprendimento con rinforzo è basato sull'analisi dei feedback. Viene spesso definito come un approccio meritocratico in quanto ricompense e punizioni incoraggiano il software a correggere e migliorare il proprio comportamento [67].

  La differenza principale con gli altri due approcci è la mancanza di un dataset per l'addestramento, che in questa soluzione non viene fornito. Al sistema viene affidato un obiettivo da raggiungere in modo incrementale: le azioni corrette che avvicinano l'agente intelligente alla soluzione vengono ricompensate e premiate, mentre vengono inflitte delle penalità quando esegue azioni che lo allontano da esso; un algoritmo di reinforcement learning, essendo a conoscenza di questo sistema, adatta il proprio comportamento per ridurre le penalità e massimizzare le ricompense [61]. Riferendosi alla definizione di machine learning, possiamo dire che il sistema, attraverso la sperimentazione e l'interazione con l'ambiente, modifica il proprio comportamento in base ai risultati ottenuti.

  Un esempio di algoritmi di apprendimento con rinforzo sono i software impiegati nelle auto a guida autonoma. Questi, infatti, grazie a un complesso sistema di sensori, sono in grado di percorrere le strade pubbliche riconoscendo eventuali ostacoli, cartelli stradali e seguendo le indicazioni; il sistema di rinforzo in questo caso funziona assegnando premi quando l'auto conduce in modo sicuro e confortevole.

2.3 Fairness

Quando si utilizzano sistemi basati sul machine learning come supporto alle decisioni, o addirittura come strumenti a cui viene delegato l'intero processo decisionale, è importante considerare le conseguenze etiche e sociali di queste scelte.

Nonostante gli algoritmi possano fornire risultati corretti dal punto di vista tecnico, potrebbero violare alcuni principi fondamentali della società: come è possibile garantire che un agente intelligente rispetti i principi di equità, di onestà, di uguaglianza o di imparzialità? Come possiamo motivare nei confronti un individuo la correttezza di una scelta presa da un algoritmo di apprendimento automatico?

Nel seguito del lavoro ci concentreremo su questi aspetti: saranno presentati due esempi specifici con l'obiettivo di comprendere meglio l'attenzione e la preoccupazione che dovremmo dedicare a questi scenari; mostreremo inoltre in che modo le nostre convinzioni, e più in generale quelle della società in cui viviamo, possono influenzare le risposte degli algoritmi; sarà mostrata la centralità dei dati durante il processo di apprendimento.

2.3.1 Esempio del COMPAS

L'utilizzo del software COMPAS evidenzia molto bene la problematica degli algoritmi di machine learning riguardo al concetto di fairness.

Negli Stati Uniti i giudici utilizzano da tempo strumenti automatici che supportano le loro decisioni; COMPAS è un'applicazione basata sull'intelligenza artificiale usata per valutare il rischio di recidiva di un detenuto sulla base di comportamenti passati di individui con le stesse caratteristiche [38]. Il software viene addestrato usando tecniche di machine learning: i dati di addestramento riguardano il comportamento dei detenuti a cui è stata concessa la libertà provvisoria e, in particolare, per ognuno di essi è indicato se è stata commessa o meno una recidiva; sulla base di queste informazioni passate il sistema assegna un grado di rischio al detenuto sotto esame [66].

Un'analisi molto interessante, e altrettanto preoccupante, è quella condotta nel 2016 dal giornale d'inchiesta ProPublica [5], che ha evidenziato un comportamento discriminatorio da parte di COMPAS. Il sistema sopravvalutava il rischio di commettere recidiva per alcune categorie di detenuti, come gli afroamericani, e questo ovviamente comportava una minore concessione della libertà.

Questo solleva importanti questioni etiche e legali riguardo alla discriminazione e alla violazione dei diritti umani.

La Dichiarazione Universale dei Diritti dell'Uomo, approvata dall'Assemblea Generale delle Nazioni Unite il 10 dicembre 1948, sottolinea all'interno dell'articolo 2 che "Ad ogni individuo spettano tutti i diritti e tutte le libertà enunciate nella presente Dichiarazione, senza distinzione alcuna, per ragioni di razza, di colore, di sesso, di lingua, di religione, di opinione politica o di altro genere, di origine nazionale o sociale, di ricchezza, di nascita o di altra condizione. Nessuna distinzione sarà inoltre stabilita sulla base dello statuto politico, giuridico o internazionale del paese o del territorio cui una persona appartiene, sia indipendente, o sottoposto ad amministrazione fiduciaria o non autonomo, o soggetto a qualsiasi limitazione di sovranità." [8].

La discriminazione basata su queste caratteristiche non dovrebbe essere tollerata, soprattutto quando si parla della vita e del futuro di una persona. Sulla base di queste considerazioni si può affermare che il comportamento del COMPAS sia stato ben lontano da quello atteso, e che rappresenta una grave violazione dei diritti umani.

2.3.2 Esempio delle assunzioni

Uno dei maggiori portali di risorse umane, HrExecutive [3], ha svolto un sondaggio tra 225 manager americani, ponendogli domande sulle modalità di reclutamento del personale. È risultato che il 66% delle grandi compagnie americane usa sistemi evoluti di intelligenza artificiale per gestire le risorse umane, e si prevede che questa salga fino all'82% entro il 2026 [30].

Con queste premesse immaginiamo uno scenario in cui un sistema basato sul machine learning sia incaricato di eseguire una prima analisi dei curriculum delle persone che si sono candidate per una determinata posizione all'interno di un'azienda. Supponiamo che i criteri sui quali si basa siano relativi a età, sesso, titolo di studio, razza ed esperienze precedenti. Dopo essere stato addestrato con curriculum di altri dipendenti, il sistema esegue la sua previsione sui nuovi candidati: alcuni vengono rifiutati mentre altri passano le prime selezioni.

Dopo un'analisi ci accorgiamo che il sistema tende a previlegiare una determinata categoria di individui, ad esempio i maschi bianchi. Nonostante questo basti a far comprendere la problematica della situazione, ancora più grave sarebbe il caso in cui un candidato volesse sapere i motivi della sua esclusione: un algoritmo di machine learning, a causa del suo non determinismo e della complessità dei modelli matematici su cui si basa, ci fornisce semplicemente un risultato, ma non dà nessuna indicazione sul processo utilizzato per ricavarlo.

2.3.3 Fairness nel machine learning

Quando si fa riferimento al concetto di fairness nel mondo del machine learning si vogliono trattare proprio i tipi di problemi descritti precedentemente.

Un punto centrale a cui prestare attenzione è il contesto in cui vengono usate tecniche di questo tipo: se il contesto non riguarda le persone allora concentrarsi esclusivamente sulla qualità degli algoritmi, la loro efficienza e la correttezza tecnica delle risposte può essere una strategia adeguata; nel momento in cui i sistemi intelligenti influenzano la vita e il futuro dei singoli individui allora entrano in gioco anche considerazioni etiche, di correttezza e di giustizia.

Un altro aspetto fondamentale, e forse anche più importante degli algoritmi stessi, su cui porre la massima attenzione sono i dati. Gli algoritmi di machine learning sono addestrati, e quindi imparano, attraverso complessi dataset, che spesso contengono informazioni private sulla vita

delle persone. Se questi contengono dei pregiudizi o delle incorrettezze allora sicuramente il sistema rifletterà quanto appreso e sarà quindi influenzato da questi problemi. Per questo motivo, quando il contesto di utilizzo del machine learning riguarda le persone, la quantità, qualità, affidabilità e completezza delle informazioni, sensibili e no, dovrebbero essere massime [64].

2.3.4 Ciclo di feedback

Le informazioni contenute nei dataset usati per l'addestramento nel machine learning, sono frutto del lavoro della raccolta dati dei nostri dispositivi; questi dati provengono quindi dall'ambiente che ci circonda e sono influenzati dalle nostre decisioni. Se le decisioni che prendiamo sono influenzate dalle attività degli agenti intelligenti, allora si instaura un ciclo continuo (Figura 4): i dati sono usati per addestrare gli algoritmi; gli algoritmi influenzano le decisioni degli umani; le persone generano nuove informazioni, cioè nuovi dati.

Questo fenomeno è ormai molto comune e in letteratura prende il nome di ciclo di feedback o amplificazione del pregiudizio [71].

Uno degli esempi più citati e comuni riportati in questo contesto è quello degli algoritmi su cui si basano i motori di ricerca [70]. Questi, una volta che gli è stata sottoposta una specifica query, generano una lista di link ordinata secondo il numero di indicizzazioni e secondo la loro tendenza.

Il classico comportamento degli utenti è selezionare uno dei primi risultati mostrati, senza spingersi nell'esplorazione di quelli che occupano posizioni inferiori nella lista. Questo comportamento genera il ciclo di feedback che abbiamo presentato precedentemente: i link in testa alla lista ottengono sempre più click e questo porta ad aumentare la loro popolarità rispetto agli altri; di ricerca in ricerca i link meno popolari partono sempre più svantaggiati e tenderanno a ricevere sempre meno attenzione.

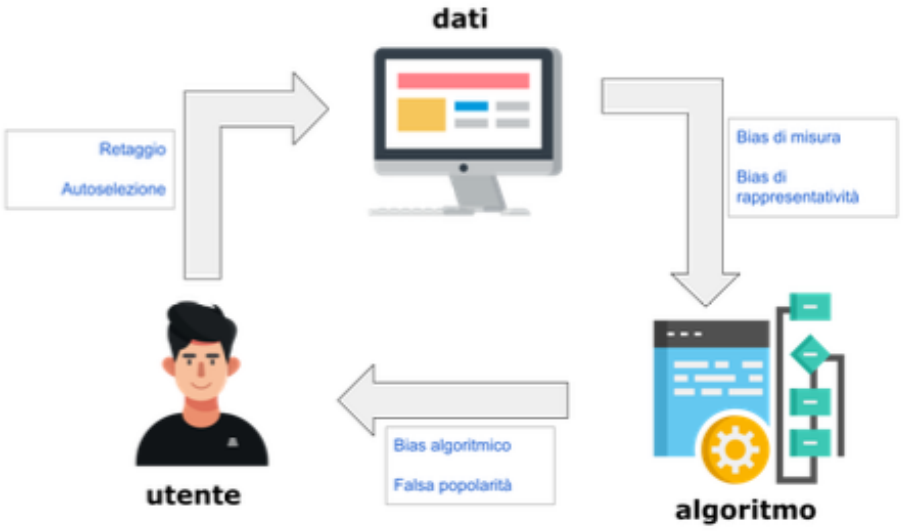

Figura 4: Ciclo di feedback [70]

2.3.5 Bias da dati ad algoritmo

Trattando il ciclo di feedback abbiamo visto che i pregiudizi, gli errori e le lacune presenti nei dati di addestramento si riflettono sulle decisioni degli algoritmi. Collegando a quanto appena detto vogliamo evidenziare ora quali sono i bias che possono compromettere la fairness dei modelli di machine learning; in particolare, presteremo attenzione a quei pregiudizi che possono essere presenti all'interno dei dataset usati per il training.

I concetti che valuteremo sono spesso sovrapposti e si possono trovare contemporaneamente all'interno di un dataset.

Il primo tipo di bias che elenchiamo è quello di misurazione, ovvero la distorsione che si manifesta a causa dello strumento, delle operazioni o dei sistemi con cui si rilevano i dati [16].

Un esempio della vita di tutti i giorni potrebbe essere un termometro non correttamente tarato al momento della rilevazione della temperatura, il che porterebbe a delle letture inaccurate; scenario simile si verificherebbe se un atleta fosse sottoposto ad una misurazione del suo peso mentre indossa degli indumenti [81, 80].

Il secondo tipo di bias che trattiamo, noto come bias di omissione, è il fenomeno che si verifica quando vengono trascurate variabili importanti per il contesto di studio. La mancanza di alcune variabili necessarie può avvenire per motivi diversi, come per mancanza di dati fondamentali, per un'impossibilità di misura oppure per un errore umano.

Ad esempio, dopo un'analisi approfondita un'azienda potrebbe far previsioni sul fatturato del prossimo trimestre; al termine del periodo indicato però si accorge che le aspettative non sono state soddisfatte. Un motivo di questo errore potrebbe non essere il metodo di analisi, ma la mancanza di informazioni, come per esempio l'assenza di informazioni sulle azioni della concorrenza [81, 80].

Il terzo tipo di pregiudizio prende il nome di bias di rappresentazione e si riferisce a quei contesti in cui il campionamento non è eseguito in modo corretto [79].

Un esempio emblematico è rappresentato dalla presenza di volti caucasici all'interno dei dataset usati per addestrare gli algoritmi di generazione di immagini. Questo particolare esempio non è ipotetico: a causa dei loro tratti caratteristici, alcuni software per il riconoscimento del volto hanno fatto fatica ad interpretare i volti asiatici; questi venivano infatti considerati come persone che continuamente sbattevano le palpebre [71].

Il quarto e ultimo tipo di bias che trattiamo è detto bias di aggregazione. Si presenta nel momento in cui si traggono conclusioni errate sugli individui perché si è osservata l'intera popolazione [70].

Ad esempio, consideriamo un insieme di immagini di cani, gatti e tigri usato per addestrare un algoritmo il cui compito è prevedere il peso di un animale che gli viene presentato. Etichettare tale insieme distinguendo solo tra "cani" e "felini" potrebbe introdurre un bias di aggregazione, visto che gatti e tigri, entrambi felini, hanno pesi moto diversi. In questo contesto, è importante che gli animali all'interno dello stesso gruppo, cioè etichettati nello stesso modo, abbiano un peso il più possibile simile [26].

3

CRITERI DI CLASSIFICAZIONE E DEFINIZIONI DI FAIRNESS

Il termine fairness, nonostante venga utilizzato anche in italiano, è una parola inglese. Se ci rivolgiamo a Google per cercare una prima definizione, il primo risultato in cui ci imbattiamo è quello del Cambridge Dictionary, che definisce la fairness come "the fact of treating everyone in the same way" e la traduce con i termini italiani equità e onestà [14, 103]. Il primo viene definito come la "Giustizia che applica la legge non rigidamente, ma temperata da umana e indulgente considerazione dei casi particolari a cui la legge si deve applicare" [18]; il secondo invece è "la qualità interiore di chi si comporta con lealtà, rettitudine e sincerità, in base a principi morali ritenuti universalmente validi" [21].

Visto che in italiano non c'è una traduzione rigorosa di fairness, se i due tentativi precedenti non ci soddisfano c'è chi propone di usare il termine uguaglianza [70], che è definita come la "condizione per cui più persone o collettività hanno diritto a essere considerate tutte alla stessa stregua, cioè pari, soprattutto nei diritti politici, sociali ed economici" [22].

Un'altra proposta che è stata portata avanti è il termine imparzialità [70]; Treccani ci rimanda al termine imparziale, che è definito come "Di persona che nel giudicare e nel trattare si mostra obiettiva e spassionata, seguendo unicamente un criterio di giustizia, senza favorire per interesse o per simpatia più gli uni che gli altri" [19].

Nonostante non ci sia una traduzione univoca del termine inglese

in esame, definire dal punto di vista filosofico e generico un concetto complesso e ambiguo come la "fairness", a prescindere che sia intesa come equità, onestà o imparzialità, non è certo lo scopo di questa tesi.

Diventa invece molto più interessante e utile definire questa idea in termini matematici, cioè cercando di stabilire una definizione univoca, non ambigua e quantificabile. Aver cercato un modo per definire un singolo concetto ed esserci imbattuti in più definizioni e idee è però un'anticipazione di quello che succede nelle situazioni simili alla nostra, cioè quando gli argomenti trattati appartengono più alle discipline umanistiche che a quelle scientifiche; una prova di quanto appena sostenuto è il numero di tentativi e di studi che sono stati svolti con lo scopo di fornire una definizione che rispetti le caratteristiche sopra elencate.

L'eliminazione dell'ambiguità nella definizione matematica della "fairness" è però un aspetto fondamentale per garantire che algoritmi e insiemi di dati possano essere valutati e confrontati in modo coerente nell'ambito della fairness. Riprendiamo la definizione di algoritmo di apprendimento: "un algoritmo di apprendimento si dice che sia in grado di imparare da un'esperienza E, con riferimento ad alcune classi di compiti T e ad alcune misure di performance P, se le sue performance nei compiti T, misurate tramite P, migliorano con l'esperienza" [77]. Nel nostro contesto, potremmo pensare di inserire il concetto di fairness all'interno dei compiti T [70]. Osservando questo possiamo affermare che, nonostante la ricerca di una definizione matematica di fairness possa essere complessa e sfidante, rappresenta un passo necessario e cruciale verso l'analisi degli algoritmi e dei sistemi automatizzati basati sul machine learning.

Nel proseguo del lavoro, saranno esposte e presentate più definizioni di "fairness", ognuna appartenente ad un criterio di classificazione specifico. È evidente come non ci sia una definizione univoca e standard di questo concetto, e come le proposte che si trovano in letteratura, nonostante siano tutte valide, offrano sfumature differenti.

I criteri di classificazione che tratteremo, cioè le modalità secondo cui verranno classificate le varie definizioni, sono quelli definiti di base, chiamati anche criteri statici di non discriminazione [36]. Lo scopo di

tali criteri è definire la presenza o l'assenza di pregiudizi sulla base di espressioni statistiche [70].

Un altro criterio sulla base del quale si possono raggruppare le definizioni di fairness, non trattato in questo lavoro ma sicuramente utile in molti contesti, è quello in base alla numerosità: in questo caso la fairness potrebbe essere misurata per l'intero gruppo sul quale si svolge l'indagine, per determinati sottogruppi, oppure per il singolo individuo [70].

3.1 Premesse

Riportiamo di seguito definizioni statistiche che possono essere utili per comprendere il contesto e le varie affermazioni trattate nel seguito:

- Probabilità statistica

  È indicata con un numero compreso tra 0 e 1 ed è il rapporto tra il numero di casi favorevoli e il numero di casi possibili.

- Unità statistica

  È un "individuo o ente sul quale viene effettuata una rilevazione statistica" [13].

- Variabile statistica

  La variabile statistica è un particolare aspetto dell'unità statistica che definisce l'oggetto dell'indagine.

- Variabile aleatoria

  Il termine aleatorio allude al fatto che il valore della variabile è determinato da un esperimento, ma il suo valore è incerto prima dell'esecuzione dello stesso [10].

- Variabile dipendente e indipendente

  La variabile indipendente, come dice il nome, è una variabile il cui valore non dipende da altre grandezze; la variabile dipendente è

invece quella il cui valore dipende da altre variabili, cioè il risultato desiderato dal ricercatore [12, 11]. Le variabili indipendenti sono usate per definire quelle dipendenti in un modello deterministico e per prevedere quelle dipendenti in un modello predittivo.

- Probabilità congiunta

  Dati due eventi A e B, $P(A \cap B)$ indica la probabilità che i due eventi si verifichino contemporaneamente.

- Probabilità condizionata

  Dati due eventi A e B, $P(A|B)$ indica la probabilità che A si verifichi sapendo che B si è verificato. Si calcola come $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

- Condizionatamente indipendenti

  Dati tre eventi A, B, C, si dice che A e B sono condizionatamente indipendenti, e si indica con $P(A \cap B|C)$, quando sapendo che C è accaduto, la conoscenza che B sia accaduto non altera la probabilità di A.

- Statisticamente indipendenti

  Due eventi A e B si dicono statisticamente indipendenti (Si indica con $A \perp B$) se e solo se, indicata con $P(A)$ e $P(B)$ la probabilità rispettivamente di A e B, $P(A \cap B) = P(A) * P(B)$.

Prima di iniziare la nostra descrizione dei vari criteri e definizioni di fairness associate, dobbiamo fare delle premesse [70, 53] sulle variabili usate in questo contesto, in modo da capire bene cosa stiamo trattando:

- La variabile che indica il gruppo di appartenenza è indicata con A; si tratta di un attributo sensibile, come ad esempio razza o sesso. È la variabile su cui è riposta la nostra attenzione: se ci fossero discriminazioni da parte del classificatore, correlate al gruppo identificato da essa, allora verrebbe meno la fairness.

- Y è la variabile dipendente, detta anche variabile target.

- $\hat{Y}$ è la variabile determinata dal classificatore.
- R è la funzione score, cioè il punteggio calcolato dal sistema basato su machine learning.
- ϵ è una quantità arbitraria non negativa e piccola.

In questo contesto, oltre a quanto precedentemente detto, ci serve capire cosa sia un classificatore e una matrice di confusione.

Pur non entrando nel dettaglio, possiamo definire informalmente un classificatore come un modello di machine learning, e quindi una funzione, che prende in input i valori delle varie caratteristiche di un oggetto e predice la classe a cui esso appartiene [87].

Una matrice di confusione, nell'ambito del machine learning anche nota come tabella di errata classificazione, rappresenta invece l'accuratezza di un classificatore. Come è mostrato in figura 5, ogni colonna rappresenta i valori predetti mentre ogni riga i valori reali; l'elemento presente nella riga $i-esima$ e colonna $j-esima$ è il numero di volte in cui il classificatore ha classificato la classe (vera) i come classe (predetta) j. Spesso la matrice risulta utile per suddividere vari casi: chiameremo veri positivi quelle situazioni in cui il classificatore restituisce $\hat{Y} = 1$ con $Y = 1$, veri negativi quando $\hat{Y} = 0$ con $Y = 0$, falsi positivi quando $\hat{Y} = 1$ con $Y = 0$ e falsi negativi quando $\hat{Y} = 0$ quando $Y = 1$ [45].

3.2 Indipendenza

Il criterio di classificazione di indipendenza è uno dei primi nati in letteratura e di conseguenza è anche uno dei più semplici. Il criterio vuole garantire che nessuno dei gruppi all'interno dell'intera popolazione dello studio sia discriminato; in questo caso l'attenzione non è focalizzata sul singolo individuo appartenente al gruppo, ma piuttosto si vuole che si abbiano risultati simili tra i gruppi presenti. Possiamo affermare, più formalmente, che il criterio di indipendenza è soddisfatto quando le

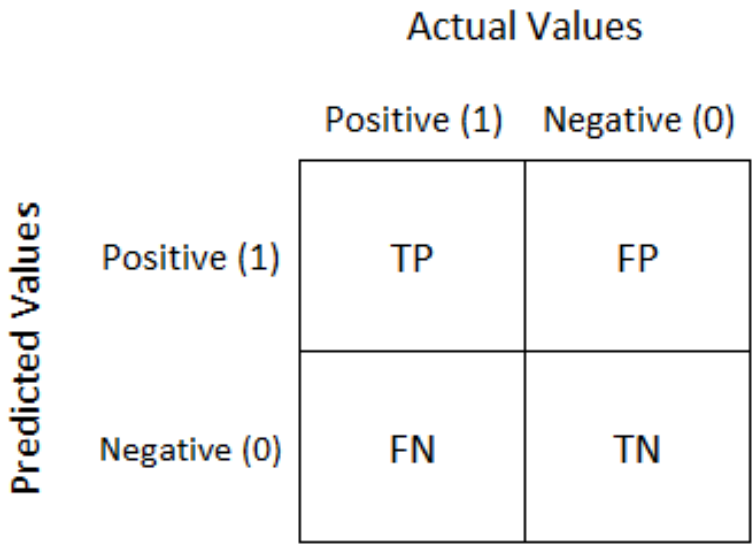

Figura 5: Matrice di confusione per classificatore binario [4]

previsioni fatte dal sistema di classificazione sono indipendenti rispetto al gruppo di appartenenza.

Di fondamentale importanza, soprattutto per l'obiettivo di questo lavoro, è osservare che tale criterio può essere impiegato anche in quelle situazioni in cui si vuole focalizzare l'attenzione esclusivamente sull'analisi oggettiva dell'equità dei dati, escludendo quindi qualsiasi distorsione o scelta del modello di machine learning utilizzato. Nelle formule che presenteremo di seguito, questa applicazione viene realizzata sostituendo $\hat{Y}$ con Y, ovvero cambiando la previsione del classificatore con il valore reale presente nei dati.

Come abbiamo anticipato nell'introduzione di questa tesi, l'obiettivo non è quello di giudicare la correttezza di un algoritmo in termini di fairness, ma piuttosto di valutare questa proprietà in relazione a un dataset usato nell'ambito del machine learning. La scelta di un criterio che separa le prestazioni di un classificatore dalle caratteristiche dei dati si allinea perfettamente con l'obiettivo che ci siamo prefissati.

3.2.1 Parità statistica

Nella definizione di parità statistica la differenza tra i due gruppi non deve essere maggiore di una certa soglia.

Formalmente questo si può esprimere come

$$|P[\hat{Y} = 1|A = 1] - P[\hat{Y} = 1|A = 0]| \leqslant \epsilon,$$

dove $A = 0$ indica l'appartenenza al gruppo di interesse, considerato svantaggiato, e con $A = 1$ l'appartenenza a un gruppo non di interesse, considerato privilegiato.

In pratica il sistema viene considerato fair se il confronto tra il gruppo di interesse e gli altri non si discosta mai di un valore non troppo grande [76].

3.2.2 Diverso impatto

In letteratura prende il nome di disparate impact, ed è simile alla precedente. Al posto di controllare la differenza di avere successo se ne considera il rapporto [76, 88].

Formalmente questo fatto può essere espresso come

$$1 - \epsilon \leqslant \frac{P[\hat{Y} = 1|A = 1]}{P[\hat{Y} = 1|A = 0]} \leqslant 1 + \epsilon.$$

3.3 Separazione

Il criterio di indipendenza non tiene conto che la variabile dipendente Y esprima un concetto di merito delle singole unità statistiche; è quindi possibile che all'interno di un gruppo ci possa essere una distribuzione di merito diversa rispetto ad altri. Il criterio di separazione ha l'obiettivo di considerare l'informazione proveniente dalla variabile target Y, cioè di considerare anche il concetto di merito e la sua distribuzione all'interno del gruppo di appartenenza della singola unità statistica.

Più formalmente il criterio di separazione è soddisfatto quando $R \perp A|Y$, cioè quando R è indipendente da A condizionatamente a Y [70, 36]. Osserviamo che in questo caso sono fondamentali le misure basate sulla matrice di confusione precedentemente introdotta: vengono sempre confrontati i valori in essa contenuti.

Questo approccio sembra essere abbastanza ragionevole. In realtà dobbiamo fare due considerazioni importanti: dobbiamo supporre che la variabile dipendente Y non sia una fonte di pregiudizi; si deve tener conto anche della veridicità delle predizioni fatte dal classificatore, cioè di $\hat{Y}$.

3.3.1 Pari opportunità

La definizione di pari opportunità può essere considerata come la versione più precisa della parità statistica, in quando tiene conto delle reali differenze presenti all'interno dei sottogruppi.

La si può formalizzare come

$$|P[\hat{Y} = 1|Y = 1, A = 1] - P[\hat{Y} = 1|Y = 1, A = 0]| \leqslant \epsilon.$$

Un sistema risulta fair secondo questa definizione se la differenza tra la probabilità di una classificazione positiva essendo veramente positiva e appartenendo al gruppo di interesse e la probabilità di una classificazione positiva essendo veramente positiva e non appartenendo al gruppo di interesse, non supera un certo ϵ fissato.

Il problema in questo caso è che la sola considerazione dei veri positivi può portare, in alcuni sistemi, a sovrastimare la proprietà di essere fair [70].

3.3.2 Quote pareggiate

Se la definizione di pari opportunità tiene conto solo dei veri positivi, con la definizione di fairness secondo le quote pareggiate si vuole integrare anche l'informazione sui falsi positivi [70].

Viene osservata sempre la differenza tra le due probabilità e la definizione può essere formalizzata come

$$|P[\hat{Y} = 1|Y = 0, A = 1] - P[\hat{Y} = 1|Y = 0, A = 0]| \leqslant \epsilon,$$

$$|P[\hat{Y} = 1|Y = 1, A = 1] - P[\hat{Y} = 1|Y = 1, A = 0]| \leqslant \epsilon.$$

In merito a questa definizione è stato osservato come, se applicata al sistema COMPAS citato nel Capitolo 2, emerga che le probabilità di essere valutati negativamente non fosse la stessa per afroamericani e caucasici; in particolare, la possibilità di ricevere la libertà era doppia per i soggetti caucasici [58, 70].

3.3.3 Parità di accuratezza complessiva

L'accuratezza di un sistema è definita come il rapporto tra la somma delle previsioni corrette (veri positivi e veri negativi) e il totale delle previsioni (veri positivi, falsi positivi, falsi negativi e veri negativi) [96]; usando le iniziali per ogni qualità presente nella matrice di confusione, questa può essere definita formalmente come

$$Acc := \frac{TP + TN}{TP + FP + FN + TN}.$$

Attraverso quanto appena detto la parità di accuratezza complessiva si può definire [70] come

$$|Acc_0 - Acc_1| \leqslant \epsilon,$$

dove Acc_0 e Acc_1 sono gli indici di accuratezza per i due gruppi rappresentati da $A = 0$ e $A = 1$, e sono definiti come

$$Acc_0 := \frac{TP_0 + TN_0}{TP_0 + FP_0 + FN_0 + TN_0},$$

$$Acc_1 := \frac{TP_1 + TN_1}{TP_1 + FP_1 + FN_1 + TN_1}.$$

e i vari termini sono

$$TP_0 = P[\hat{Y} = 1|Y = 1, A = 0],$$

$$TN_0 = P[\hat{Y} = 0|Y = 0, A = 0],$$

$$FP_0 = P[\hat{Y} = 1|Y = 0, A = 0],$$

$$FN_0 = P[\hat{Y} = 0|Y = 1, A = 0],$$

$$TP_1 = P[\hat{Y} = 1|Y = 1, A = 1],$$

$$TN_1 = P[\hat{Y} = 0|Y = 0, A = 1],$$

$$FP_1 = P[\hat{Y} = 1|Y = 0, A = 1],$$

$$FN_1 = P[\hat{Y} = 0|Y = 1, A = 1].$$

3.3.4 Parità di trattamento

La proprietà di fairness è soddisfatta dalla parità di trattamento se il rapporto tra falsi positivi e falsi negativi nei diversi gruppi non si discosta da un valore fissato [70].

Questa può essere definita formalmente come

$$|\frac{P[\hat{Y} = 1|Y = 0, A = 0]}{P[\hat{Y} = 0|Y = 1, A = 0]} - \frac{P[\hat{Y} = 1|Y = 0, A = 1]}{P[\hat{Y} = 0|Y = 1, A = 1]}| \leqslant \epsilon.$$

3.4 Sufficienza

Le definizioni di fairness che rispettano il criterio di sufficienza sono anche dette basate sulla calibrazione. Questo concetto permette di ottenere una misura quantitativa della probabilità di ricevere dal sistema un certo punteggio; nella nostra situazione avremo che per ogni valore assunto da $R = r$ si ha $r \in [0, 1]$. Se consideriamo le singole unità statistiche la calibrazione può essere espressa indipendentemente dal gruppo di appartenenza come $P(Y = 1|R = r) = r$. In generale il criterio di sufficienza è

soddisfatto quando $Y \perp A|R$, cioè quando Y è statisticamente indipendente dalla variabile A condizionatamente al punteggio R [70].

L'obiettivo del criterio è quello di assicurare che il modello fornisca i risultati desiderati per tutte le sotto popolazioni, cioè si richiede non solo che funzioni bene con tutta la popolazione nel suo complesso, ma anche per gruppi specifici all'interno di essa [44].

3.4.1 Calibrazione

Il principio di calibrazione stabilisce che a parità di score si abbiano esiti uguali, indipendentemente dal gruppo di appartenenza.

Formalmente la si può esprimere come

$$|P[Y = 1|R = r, A = 1] - P[Y = 1|R = r, A = 0]| \leqslant \epsilon.$$

Questa definizione è simile a quella di pari opportunità, con la differenza che in questo caso si tiene conto anche del punteggio ottenuto dalla singola unità statistica [70].

3.4.2 Buona calibrazione

La definizione di buona calibrazione è simile alla precedente, ma in questo caso la probabilità di risultare positivo non si deve discostare dallo score stesso.

Formalmente si può definire come

$$|P[Y = 1|R = r, A = 1] - r| \leqslant \epsilon,$$

$$|P[Y = 1|R = r, A = 0] - r| \leqslant \epsilon.$$

In questo caso si ottiene lo stesso vantaggio della calibrazione, cioè che la probabilità di una valutazione positiva sia simile tra i diversi gruppi, ma in più si impone che per tutti i gruppi tale probabilità non si discosti dalla probabilità vera di r [70].

4

DATASET PER LA FAIRNESS

Nel Capitolo 3 abbiamo provato a fornire una traduzione del termine inglese "fairness", ottenendo però risultati non del tutto soddisfacenti a causa dell'ambiguità e della vasta gamma di soluzioni proposte. Per affrontare questo problema in un contesto a noi più familiare, abbiamo introdotto delle definizioni formali, incentrate su concetti matematici. Questo approccio ci ha permesso di eliminare le ambiguità incontrate precedentemente e di fornire gli strumenti per un'analisi più accurata e approfondita.

Sempre relativamente al concetto di fairness, nel Capitolo 2 abbiamo presentato alcuni esempi della sua violazione, con particolare enfasi su uno dei casi più significativi: quello del COMPAS è infatti una delle situazioni più trattate in letteratura in questo ambito, a causa delle questioni relative alla libertà degli individui.

Successivamente è stato presentato il ciclo di feedback e le modalità con cui i pregiudizi possono essere introdotti in modo ciclico all'interno dei sistemi basati sul machine learning. In questo ambito è stato approfondito il bias da dato ad algoritmo, il quale viene introdotto nel processo quando il sistema è addestrato con dati che contengono dei bias.

Il lavoro portato avanti fino a questo punto mette in evidenza la centralità dei dati nell'AI e in particolare nella pratica del machine learning. Senza una vasta quantità di informazioni, infatti, gli algoritmi basati sul machine learning avrebbero una capacità limitata di fare previsioni o di prendere decisioni accurate.

Oltre a fornire le basi necessarie per la costruzione dei modelli usati nelle aziende e organizzazioni, i dataset rivestono un ruolo fondamentale anche nella ricerca accademica, consentendo di applicare e testare le soluzioni di machine learning sotto esame in contesti realisti e veritieri. In questa situazione selezionare il dataset più appropriato per il proprio dominio di interesse non è però banale. Sul web si trovano molti dataset relativi alla fairness, ognuno dei quali copre un campo applicativo diverso: ci sono dataset che appartengono al settore dell'industria, della sanità, dell'istruzione o delle assunzioni. Inoltre, ciascuno di questi può essere contestualizzato in modi diversi e può essere applicato a definizioni diverse di fairness [66].

In questo contesto, è stato trovato un lavoro che ha avuto un ruolo significativo nella comprensione di quanto appena discusso; questo può essere altrettanto fondamentale per chiunque stia cercando un dataset da usare come benchmark. Lo studio in questione è intitolato "A survey on datasets for fairness-aware machine learning" [66] ed è stato pubblicato nel 2022 dagli autori Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang ed Eirini Ntoutsi.

Si tratta di una raccolta esaustiva dei dataset più utilizzati e citati all'interno del web in ambito fairness. Ogni dataset è stato analizzato accuratamente ed è descritto in modo approfondito; sono elencate infatti le sue caratteristiche di base come cardinalità, dimensione dimensione e gli attributi sensibili tipicamente considerati nella ricerca.

4.1 Attributi sensibili

I dataset che riguardano la fairness di maggior interesse sono ovviamente quelli che contengono informazioni personali sugli individui; è ragionevole affermare che, in queste situazioni, alcune delle caratteristiche che descrivono le varie istanze siano attributi sensibili. Se provassimo a stilare una lista di attributi di questo tipo, è probabile che la maggior parte di noi includerebbe senza molti dubbi la razza e il genere. Pur-

troppo, o per fortuna, la strada che porta ad avere una lista esaustiva e dettagliata degli attributi sensibili è molto complessa. Questo problema riveste un ruolo di così spiccato interesse, che un punto di partenza per capire quali attributi siano sensibili e quali no ci è fornito direttamente dai più importanti regolamenti emanati dall'Unione Europea.

Nel "Regolamento per la protezione dei dati personali" (GDPR, "General Data Protection Regulation") definito dall'Unione Europea nel 2016 [17], i dati sensibili vengono definiti, con riferimento al punto 51 dell'articolo, come quelle informazioni 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" [38].

In ambito privacy, lo stesso GDPR definisce molte regole per la protezione dei dati sensibili. Un esempio che possiamo citare in merito è il segreto professionale dei medici: quando viene scoperta una malattia, al medico curante è consentito trasmettere l'informazione solo ai soggetti che sono coinvolti nella cura del paziente, come ad esempio altri medici o infermieri. Altri esempi che possiamo portare riguardano le situazioni presenti all'interno di alcuni processi decisionali: durante un concorso per la assunzioni di nuovi lavoratori si tende a omettere le informazioni personali dei candidati, come razza, genere, nome e cognome, in modo da garantire imparzialità ed evitare favoreggiamenti [38].

Le informazioni sensibili rivestono un ruolo fondamentale anche nell'ambito della fairness nel machine learning. Infatti, quando la classificazione effettuata da un sistema di questo tipo produce un risultato che mostra una correlazione con uno o più attributi sensibili, essa è ritenuta ingiusta e non accettabile a livello sociale ed etico.

In questa situazione sarebbe necessario introdurre delle correzioni nei dati o all'interno del sistema stesso, con lo scopo di eliminarla totalmente o quanto meno ridurla [70]. Questi tipi di cambiamenti devono essere però applicati con molta attenzione: escludere informazioni può in alcuni

casi essere una pratica consigliata, ma dobbiamo essere consapevoli che in altri può invece portare al bias di omissione trattato nel Capitolo 2 [103].

4.2 Adult

Nonostante la presenza di una grande disponibilità di dataset, è interessante osservare che la maggior parte dei lavori di ricerca sulla fairness, specialmente quando si tratta di analizzare ed utilizzare dati strutturati sotto forma di tabelle, coinvolge un insieme non così vasto di dataset rispetto a quelli presenti.

L'UCI Machine Learning Repository, una raccolta di database e generatori di dati creata dall'iniziativa di un dottorando nel 1987 e ancora oggi molto utilizzata [1], ha messo a disposizione di ricercatori, studenti e educatori una vasta gamma di dataset da utilizzare sia nell'ambito industriale che in quello accademico.

Nel 2021 il secondo dataset più popolare in questo ambito era uno di quello presenti in questa raccolta, l'UCI Adult. Usato per la prima volta nel 2009 e comparso in più di 300 articoli di ricerca sulla fairness, è un dataset derivato da un censimento del 1994, condotto dal Census Bureau degli Stati Uniti. È interessante notare come ci sia un dataset identico chiamato "Census Income Data Set" e un dataset più grande, strettamente correlato, chiamato "Census-Income (KDD) Data Set" [51].

Il compito di classificazione del sistema di machine learning che utilizza questo dataset, consiste nel decidere se il reddito annuale di una persona supera i 50000 dollari statunitensi sulla base delle caratteristiche demografiche dell'individuo.

Il dataset è composto da 48842 istanze, ognuna composta da 15 attributi di cui 6 sono numerici, 7 categorici e 2 binari; nella figura 6 viene mostrata una panoramica più approfondita di quanto appena descritto. Nonostante la validità complessiva delle informazioni presenti, è importante precisare

Attributes	Type	Values	#Missing values	Description
age	Numerical	[17–90]	0	The age of an individual
workclass	Categorical	7	2,799	The employment status (private, state-gov, etc.)
fnlwgt	Numerical	[13,492–1,490,400]	0	The final weight
education	Categorical	16	0	The highest level of education
educational-num	Numerical	1–16	0	The highest level of education achieved in numerical form
marital-status	Categorical	7	0	The marital status
occupation	Categorical	14	2,809	The general type of occupation
relationship	Categorical	6	0	Represents what this individual is relative to others
race	Categorical	5	0	Race
sex	Binary	{Male, Female}	9	The biological sex of the individual
capital-gain	Numerical	[0–99,999]	0	The capital gains for an individual
capital-loss	Numerical	[0–4,356]	0	The capital loss for an individual
hours-per-week	Numerical	[1–99]	0	The hours an individual has reported to work per week
native-country	Categorical	41	857	The country of origin for an individual
income	Binary	{≤50K, >50K}	0	Whether or not an individual makes more than $50,000 annually

Figura 6: Caratteristiche degli attributi di Adult [66]

che per 3620 (7,41%) record abbiamo dei valori mancanti [66].

In letteratura, quando si utilizza questo dataset per gli studi sulla fairness, solitamente i seguenti attributi sono indicati come fattori scatenanti di bias:

- $sesso = \{male, female\}$

  Il genere maschile è quello predominante rispetto a quello femminile: si parla di circa 32500 (circa il 67%) di istanze di sesso maschile contro circa 16000 di sesso femminile.

- $race = \{white, black, asian - pac - islander, amer - indian - eskimo, other\}$

  Solitamente la razza è indicata in modo binario come $race = \{white, non - white\}$. Il dataset è dominato da persone bianche: il rapporto è di circa 6:1, approssimativamente l'85% contro il 15%.

- $età = [17, 90]$

  Solitamente non viene utilizzato un valore discreto per l'età, ma piuttosto si categorizza in tre intervalli: minore di 25, tra 25 e 60 e maggiore di 60. La maggior parte delle istanze del dataset appartengono al secondo gruppo, quello tra i 25 e i 60 anni, con circa 35000 istanze, per una percentuale del 77% circa.

Osserviamo che in alcuni lavori in cui è stato analizzato questo dataset sono stati considerati come dati sensibili anche lo stato civile e il paese di origine [38].

5

APPLICAZIONE DEFINIZIONI AL DATASET ADULT

Il fine ultimo di questo Capitolo è valutare se un dataset soddisfa la proprietà di fairness secondo alcune delle definizioni esaminate e descritte nel Capitolo 3. In particolare, verrà analizzato il dataset Adult alla luce delle definizioni che soddisfano il criterio di indipendenza. Gli attributi verso i quali si vuole verificare l'assenza o la presenza di bias sono due: la razza (*race*) e il sesso (*sex*).

Tale analisi ha l'obiettivo di determinare se definizioni diverse conducano allo stesso o a risultati diversi, variando inoltre anche l'attributo sensibile considerato. È importante sottolineare come questa analisi non sia completa ed esaustiva, e non possa portare a conclusioni generali; si tratta in ogni caso di un'applicazione delle conoscenze acquisite fino a questo momento.

Tra tutti i dataset disponibili all'interno del lavoro menzionato ad inizio del Capitolo 4, la scelta di Adult come oggetto di studio è stata dettata dal suo vasto impiego in questo ambito. Esso è infatti molto conosciuto e comprensibile, il che lo rende un ottimo candidato per il nostro lavoro.

La decisione di concentrarsi sulle definizioni di parità statistica e diverso impatto verte invece sulla considerazione che l'obiettivo di questa tesi non è focalizzato sugli algoritmi di machine learning e sul loro utilizzo. Implementare definizioni che soddisfano gli altri criteri, cioè quelli di separazione e sufficienza, richiederebbe l'applicazione al dataset di un algoritmo di classificazione. Ciò potrebbe portare a problemi non banali e, soprattutto, non pertinenti con l'obiettivo di questa tesi.

Nel seguito, prima di esporre il metodo utilizzato per giungere ai risultati, verranno brevemente introdotti Python e Pandas, strumenti utilizzati in questa fase del lavoro. L'intento non è ovviamente quello di fornire una conoscenza approfondita ed esaustiva di tali software, ma piuttosto si vuole offrire al lettore una base di partenza nel caso siano argomenti mai trattati.

5.1 Python

Nel vasto e complesso mondo dei linguaggi di programmazione, Python, linguaggio general-purpose di alto livello e open source, assume oggi giorno una posizione di notevole rilievo. La sua diffusione e il suo ampio impiego non sono evidenti solamente agli specialisti del settore e ai professionisti che operano in ambito aziendale, ma emergono chiaramente anche a chi è alle prime armi e non ha molta esperienza. Osservando le domande poste sulla piattaforma Stack Overflow, è interessante notare come più del 15% siano taggate "python" [90], una percentuale che è in crescita costante e che sicuramente non può essere ignorata.

Individuare i motivi per cui un linguaggio di programmazione ha più successo rispetto ad un altro non è sicuramente un compito facile. Il nostro obiettivo non è infatti descrivere in modo accurato ogni caratteristica di Python, ma cercheremo piuttosto di elencare alcune delle sue particolarità che lo hanno reso così popolare.

Innanzitutto, collochiamolo temporalmente: è nato ufficialmente nel 1991, quando Guido Van Rossum ha rilasciato la sua prima versione, la 0.9.0 [28]. Questo lungo periodo non solo ha concesso al linguaggio il tempo necessario per crescere e maturare, ma ha anche favorito lo sviluppo di una vasta e solida comunità di sviluppatori pronta ad offrire supporto e assistenza ai più inesperti in caso di problemi o difficoltà.

Una delle caratteristiche distintive di Python è la sua estrema flessibilità. Nonostante sia comunemente categorizzato come un linguaggio object-oriented, permette di sviluppare codice seguendo anche altri paradigmi

di programmazione, come quello imperativo e funzionale [56]. Questo permette agli sviluppatori di adottare l'approccio che più desiderano in base al contesto.

Un altro punto di forza di Python è la sua semplicità, tanto che da alcuni viene paragonato ad uno pseudo-linguaggio [57]. La sintassi è decisamente meno complessa rispetto agli altri linguaggi di programmazione di pari livello e adozione, come Java o C++. In Python le parentesi graffe sono sostituite dall'indentazione, aspetto che, nonostante possano rendere il codice meno manutenibile in grandi progetti, lo rendono sicuramente leggibile in modo più intuitivo.

Anche Python ha degli aspetti negativi e un rovescio della medaglia. Due caratteristiche da tenere in considerazione quando lo si utilizza per lo sviluppo sono il suo essere un linguaggio interpretato e il fatto di basarsi su una tipizzazione dinamica. [90, 52].

Questa proprietà implica non avere un controllo statico sul tipo degli oggetti con cui si lavora, il che può portare ad una perdita di affidabilità aumentando il numero degli errori a runtime: infatti, è il solo programmatore che controlla, ad esempio, la correttezza degli assegnamenti e il passaggio adeguato degli argomenti alle funzioni.

Altra conseguenza dell'essere un linguaggio interpretato è la perdita di efficienza. La traduzione in istruzioni macchina eseguibili dal calcolatore viene svolta durante l'esecuzione del programma, e questo richiede tempo aggiuntivo. Inoltre, il codice non può essere ottimizzato e verrà sempre eseguito come il programmatore lo ha pensato: il traduttore, infatti, non ha una visione globale del programma, ma solo della singola istruzione che interpreta ed esegue.

5.1.1 Modularità

Prima di procede oltre dobbiamo considerazione una caratteristica fondamentale di Python, la sua modularità. Questo linguaggio non possiede nativamente le classi necessarie per eseguire computazioni avanzate e complesse, ma la sua potenza si basa sulla grande disponibilità di librerie

esterne open source, accessibili e utilizzabili da chiunque.

Tra più popolari che possiamo citare, quelle che contribuiscono a renderlo un linguaggio semplice ma adatto ad una vasta gamma di applicazioni, spiccano NumPy per l'elaborazione numerica [32], Pandas per l'analisi dei dati [25] e Scikit-Learn per il machine learning [34]. Le prime due appartengono a SciPy, la principale suite del noto "Python Scientific Stack", un set di librerie open source sponsorizzato da PyData, che ne promuove lo sviluppo anche da punto di vista economico [91].

La comunità di sviluppatori Python può usufruire del "Python Package Index" (PyPI). Si tratta di un vasto archivio di software contenente i lavori e i progetti di altri sviluppatori o organizzazioni [33]. I moduli presenti all'interno di questa collezione vengono istallati tramite il terminale utilizzando il comando *pip*.

Per procedere ci sono due soluzioni: installare il pacchetto a livello di sistema o all'interno di un ambiente virtuale. La prima delle due è sconsigliabile per diversi motivi. L'aspetto principale è che progetti Python diversi possono richiedere pacchetti diversi o versioni diverse dello stesso pacchetto, il che potrebbe causare conflitti se tutto fosse installato nella stessa directory.

È pertanto consigliato usare un ambiente virtuale distinto per ciascun progetto Python. Un ambiente virtuale costituisce uno spazio isolato e indipendente dal resto del sistema; concretamente si tratta di una cartella contenente i file necessari per il funzionamento dell'ambiente, una copia dei file binari di Python e tutti i moduli installati e che vogliamo utilizzare nel progetto in questione. Il metodo nativo offerto da Python per creare un ambiente virtuale è il comando *venv*; un'alternativa valida e decisamente più comoda è il comando *conda* messo a disposizione da Anaconda, tool che offre diverse funzionalità aggiuntive.

5.2 Pandas

Pandas è una libreria open source scritta per il linguaggio di programmazione Python. Il nome deriva da "panel data", un termine comune in ambito statistico per indicare insiemi di dati che contengono misure ripetute della stessa variabile [40, 75].

Il progetto è nato nel 2008 da Wes McKinney; l'obiettivo dello sviluppatore statunitense era quello di colmare il divario qualitativo tra gli strumenti per l'analisi dei dati in Python e le numerose piattaforme specifiche di questo dominio, come R e MATLAB. Originariamente ideata come un tool per l'analisi dei dati finanziari, Pandas si è evoluto nel corso del tempo, diventano oggi una delle librerie maggiormente utilizzate per condurre analisi di dati in modo semplice e veloce [75].

Nel paragrafo precedente abbiamo discusso del fatto che Python non sia un linguaggio di programmazione eccellente dal punto di vista delle prestazioni. Per ottimizzare la velocità di elaborazione, Pandas si basa sulla già citata libreria NumPy, che permette di processare array monodimensionali e multidimensionali in modo estremamente rapido [32].

Questo approccio, tuttavia, nonostante renda Pandas uno strumento molto efficiente, può essere considerato anche il suo punto debole. Le strutture dati messe a disposizione sono infatti progettate per eseguire analisi di dati in-memory. Di conseguenza questo rende Pandas uno strumento non adatto quando la dimensione dei dati superano la capacità della memoria [89].

5.2.1 Strutture dati

Le funzionalità che Pandas mette a disposizione si basano su due strutture dati principali, le series e i dataframe [89]. Queste coprono praticamente tutte le esigenze di rappresentazione dei dati usati in ambito finanziario, statistico, scientifico e ingegneristico.

Una serie è sostanzialmente un array unidimensionale labeled, cioè

in cui ogni elemento è identificato non univocamente da un'etichetta. Nonostante convenzionalmente le etichette siano numeri naturali che iniziano da 0, possono anche essere specificate all'interno di un array con la stessa cardinalità della serie [89]. È bene specificare come una struttura di questo tipo abbia la capacità di contenere oggetti di tipo eterogeneo (interi, stringhe o oggetti Python).

In modo simile ad una serie, un dataframe è un array bidimensionale labeled, in cui colonne diverse possono essere popolate da oggetti di tipo diverso; praticamente si può pensare come un dizionario di serie, cioè una struttura dati dove la chiave non deve essere necessariamente un intero e il valore è dato da un insieme di elementi di tipo eterogeneo [52]. L'obiettivo dei dataframe di Pandas è risolvere il principale problema degli array strutturati di NumPy: vogliono offrire la flessibilità necessaria per lavorare con array di tipo eterogeneo, come avviene negli altri ambienti di lavoro statistici [75, 89].

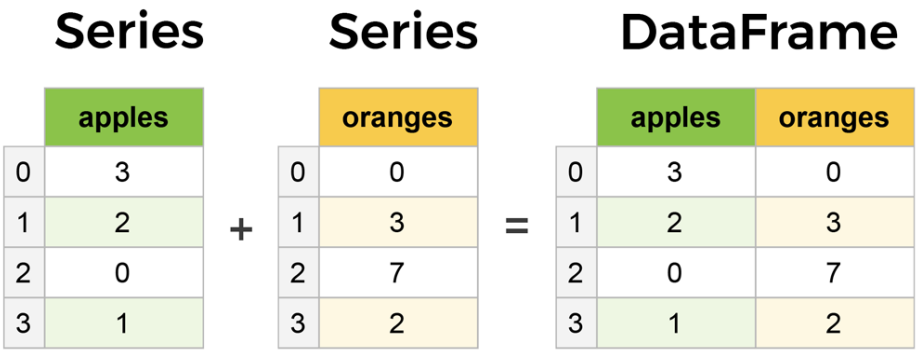

Figura 7: Serie e dataframe di Pandas [6]

5.2.2 Funzionalità principali

Basandosi su quanto introdotto nel paragrafo precedente vogliamo elencare e descrivere brevemente le funzionalità principali presenti in Pandas. Il tool permette, tra le altre, le seguenti operazioni [91, 86]:

- Gestione dai dati mancanti, rappresentati con *NaN*, per ogni tipo.

- Funzioni di raggruppamento, divisione e combinazione tra dataset.
- Funzioni di indicizzazione, anche gerarchica, tramite label.
- Funzioni per reperire informazioni dai dati.
- Allineamento intelligente dei dati.
- Funzioni di I/O per interagire con dati in formati diversi, tra i quali CSV, Excel e database SQL.
- Funzioni per modificare la dimensione del dataframe, aggiungendo o rimuovendo colonne.

5.3 Analisi

Per agevolare il lettore riportiamo di seguito le definizioni di parità statistica e diverso impatto: un sistema rispetta la fairness secondo la definizione di parità statistica quando $|P[\bar{Y} = 1|A = 1] - P[\bar{Y} = 1|A = 0]| \leqslant \epsilon$; un sistema rispetta la fairness secondo la definizione di diverso impatto quando $1 - \epsilon \leqslant \frac{P[\bar{Y} = 1|A = 1]}{P[\bar{Y} = 1|A = 0]} \leqslant 1 + \epsilon$.

L'obiettivo è quindi garantire che la probabilità di successo (stipendio superiore a 50 mila dollari), data l'appartenenza al gruppo (uomini/donne e bianchi/non bianchi), sia la stessa per entrambi i gruppi. Nel caso della parità statistica questo sarà controllato attraverso la differenza; nel caso del diverso impatto attraverso il rapporto.

Nel contesto del nostro studio, con l'espressione $\bar{Y} = 1$ si intende la situazione in cui il soggetto relativo all'istanza in questione abbia uno stipendio maggiore di 50000\$; con l'espressione $A = 1$ si intende invece l'appartenenza del record alla classe di sesso $Female$ in un caso e alla razza $Non - white$ nell'altro.

Listing 1: Operazioni preliminari

```
#lettura file
header = ["age", "workclass", "fnlwgt", "education",
    "education-num", "marital-status", "occupation",
    "relationship", "race", "sex", "capital-gain", "capital-loss",
    "hours-per-week", "native-country", "income"]
dataframe = pd.read_csv("/Users/edoardosarri/Library/Mobile
    Documents/com~apple~CloudDocs/UniFi/Tesi/Mia/Codice/adult.csv",
    header=None, names=header)

#operazioni preliminari
dataframe.replace("?", np.nan, inplace=True)
dataframe.dropna(inplace=True)
dataframe.loc[dataframe["race"] != "White", "race"] = "Non-white"
```

5.3.1 Operazioni preliminari

Il dataset, scaricato da GitHub [23], non è stato utilizzato direttamente nella sua forma originale. Come si può osservare dal codice Python 1, sono state svolte le seguenti modifiche:

- Poiché il file CSV non include un header, cioè le etichette che identificano le colonne, è stato aggiunto manualmente in fase di lettura.
- Come descritto durante la presentazione del dataset, Adult contiene 3620 record con valori mancanti, rappresentati nel file scaricato dalla stringa "?". Poiché l'eliminazione del 7% delle istanze totali non influisce significativamente sui risultati della nostra analisi, tali record sono stati eliminati dal dataframe.
- All'interno del dataset, i valori dell'attributo $Race$ non sono binari. Per i nostri scopi, seguendo anche quanto fatto in molti altri studi, sono necessari solo i valori $White$ e $Non-white$. Per questo motivo tutti i valori diversi da $White$ sono rimpiazzati con $Non-white$.

Listing 2: Script

```
#genere
dfFemale = filtra(dataframe,"sex","Female")
dfFemaleIncome = filtraIncome(dfFemale)
totFemale = dfFemale.shape[0]
totFemaleIncome = dfFemaleIncome.shape[0]
probFemale = totFemaleIncome / totFemale

dfMale = filtra(dataframe,"sex","Male")
dfMaleIncome = filtraIncome(dfMale)
totMale = dfMale.shape[0]
totMaleIncome = dfMaleIncome.shape[0]
probMale = totMaleIncome / totMale

paritaStatisticaGenere = abs(probFemale-probMale)
diversoImpattoGenere = probFemale/probMale
```

5.3.2 Risultati

L'esecuzione del codice 2 ha prodotto risultati sicuramente interessanti, ma allo stesso tempo prevedibili in base alle esperienze di vita quotidiana che ognuno di noi porta con sé. Come le nostre misure e analisi successivamente mostreranno, è chiaramente evidente che il dataset Adult non rispetta la fairness per nessun dei due attributi sensibili considerati, ma in particolar modo per quanto riguarda il genere.

È importante sottolineare che il codice presentato in questa circostanza non è completo: è stata omessa la parte relativa alla generazione dei file PDF con i risultati, in quanto banale e non interessante in questo contesto; è inoltre illustrato solamente il codice relativo all'analisi sul genere, poiché che quello relativo alla razza segue lo stesso principio e la stessa logica.

GENERE

- Istanze di genere femminile: 14695
- Istanze di genere femminile con income >50K$: 1669
- Probabilità per il genere femminile di avere un income >50K$: 0.11357604627424293

- Istanze di genere maschile: 30527
- Istanze di genere maschile con income >50K$: 9539
- Probabilità per il genere maschile di avere un income >50K$: 0.31247747895305794

- Parità statistica: 0.198901432678815
- Diverso impatto: 0.3634695423643793

Figura 8: Risultati genere

Per quanto riguarda il genere, come si può osservare dalla figura 8, le istanze di genere femminile che percepiscono un reddito maggiore di 50 mila dollari sono 1669 su un totale di 14695; ciò implica che la probabilità di avere uno stipendio maggiore a 50 mila dollari, è appena superiore all'11%.

Una situazione nettamente diversa è quella che si riscontra per il genere maschile: con 9539 record su un totale di 30527, per l'uomo la probabilità di avere una retribuzione maggiore di 50 mila dollari è superiore al 31%. Si può quindi concludere che il genere maschile è circa tre volte più privilegiato rispetto a quello femminile.

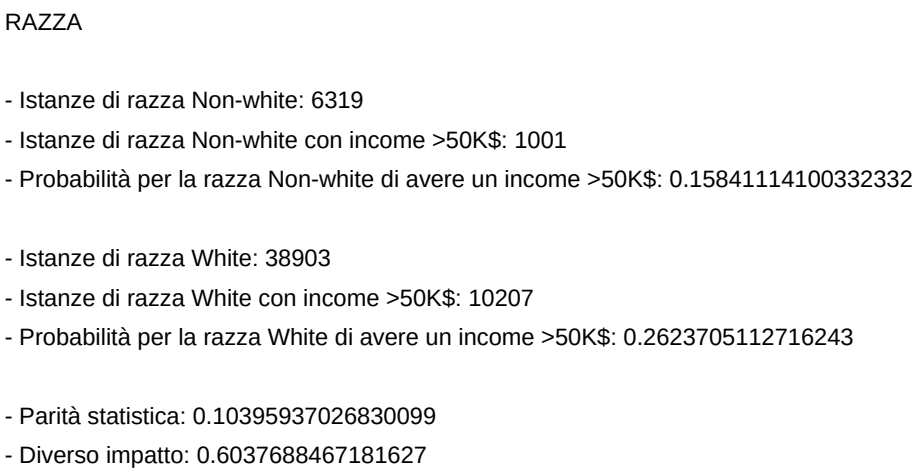

Figura 9: Risultati razza

Discorso analogo si può fare per la razza. Come mostrato in figura 9, la probabilità per una persona di razza "Non bianca" di percepire uno stipendio superiore ai 50 mila dollari è di poco superiore al 15%; per una persona di razza "Bianca" la probabilità invece sale poco al di sopra del 26%.

Rispetto allo scenario osservato per il genere, in questo cosa la situazione è migliore: sebbene ci sia comunque uno sbilanciamento a favore della razza "Bianca", si può notare che questo è di poco superiore al 60%, un panorama nettamente migliore rispetto al precedente.

6

CONCLUSIONI

I dati costituiscono una perfetta rappresentazione del mondo, trasformando ogni aspetto della complessa realtà in cui tutti noi viviamo in forma digitale. In questo contesto di pervasività dei dati all'interno della vita quotidiana, al giorno d'oggi assistiamo ad una progressiva fusione del mondo analogico con quello digitale. Questa interconnessione, che continua ad evolversi ad un ritmo sempre più elevato, rende sempre più difficile le sfide che ricercatori e professionisti devono affrontare nello stabilire una relazione tra le regole etiche e le nuove tecnologie in via di sviluppo.

Nell'ampio panorama del machine learning, la questione della fairness emerge come un aspetto fondamentale e nel futuro è destinata a ricevere sempre più attenzioni. Si tratta infatti di un argomento cruciale, visto che in questa disciplina essa incide direttamente sull'imparzialità dei risultati ottenuti dall'esecuzione di algoritmi di apprendimento automatico. La necessità di risolvere questo problema, o quanto meno di definirlo in un modo chiaro e univoco, deriva dalla consapevolezza che la presenza di pregiudizi, all'interno di tali modelli, può influenzare in modo significativo la vita delle persone in diversi ambiti.

In questa tesi di laurea, ci siamo focalizzati sui problemi relativi alla presenza di bias all'interno dei dataset impiegati per l'addestramento dei modelli e degli algoritmi di machine learning. Si è cercato di capire come queste discriminazioni possano influenzare la vita delle persone e, successivamente, abbiamo posto l'attenzione sulle capacità attuali di identificare in modo matematico la presenza di queste problematiche.

In primo luogo, all'inizio del Capitolo 2, abbiamo introdotto la storia e il funzionamento dell'intelligenza artificiale e del machine learning. È poi stato mostrato come queste tecnologie possano influenzare in modo negativo la vita delle persone in molte situazioni.

I due esempi discussi, quello del COMPAS e delle assunzioni, pur essendo diversi per natura e contesto, ci portano entrambi a porre l'attenzione sul preoccupante utilizzo di questi sistemi decisionali automatici nei contesti dove è in gioco il futuro degli individui. Il loro utilizzo in situazioni simili a quelle della concessione della libertà e dell'opportunità lavorativa, richiedono un'attenta analisi e supervisione al fine di evitare ogni forma di discriminazione, ingiustizia e pregiudizio.

Un ulteriore aspetto su cui ci siamo concentrati, e che abbiamo sottolineato nella parte finale del Capitolo 2, è la stretta correlazione tra i dati di training, utilizzati per l'addestramento dei modelli di machine learning, e le risposte che tali algoritmi generano.

Nonostante possa sembrare intuitivo credere che gli algoritmi di machine learning siano addestrati esclusivamente con dati provenienti dall'ambiente e totalmente disgiunti da esso, abbiamo mostrato che tali informazioni possono in realtà rappresentare decisioni prese dall'uomo sulla base delle risposte fornite dagli algoritmi stessi.

In altre parole, se i sistemi producono risposte che contengono bias o distorsioni, i dati rifletteranno a loro volta tali pregiudizi e in questo modo si crea un ciclo di condizionamento, che non fa altro che amplificare il problema.

Nel Capitolo 3, abbiamo osservato che le traduzioni italiane di fairness sono molte e molto diverse tra loro. Nel seguito abbiamo esposto le definizioni formali e matematiche di questo concetto che al momento si trovano in letteratura; abbiamo visto che ognuna di queste ricade all'interno di uno dei principi di indipendenza, separazione o sufficienza.

Questa parte di percorso ci ha consentito di comprendere la difficoltà di definire un complesso concetto umanistico come la fairness in ambito matematico, e quindi di tutti i problemi relativi alla trasformazione di

una definizione qualitativa in una quantitativa. Abbiamo infatti notato come non ci sia misura standard e univoca, bensì una molteplicità di approcci che variano tra loro. La problematica principale da affrontare è quella di definire la metrica da utilizzare, cercando di raggiungere il giusto bilanciamento tra accuratezza ed equità [70].

La mancanza di un'unica definizione comporta considerevoli problemi quando si tratta di migliorare la fairness nei dataset e, di conseguenza, negli algoritmi di machine learning. Infatti, vista la grande differenza tra i vari criteri, adottare strategie il cui obiettivo è migliorare questa proprietà in ogni suo aspetto non sembra essere una strategia utilizzabile.

Dopo aver presentato, all'interno del Capitolo 4, alcuni dataset per la fairness e le loro caratteristiche, nel capito 5 è stata condotta un'analisi dettagliata sul dataset Adult.

È stato mostrato come, nonostante sia uno dei dataset più utilizzati in letteratura, non rispetta le due definizioni che soddisfano il criterio di indipendenza. In particolare, si è evidenziato uno sbilanciamento significativo del dataset a favore degli uomini se consideriamo il genere come attributo sensibile, e dei bianchi se consideriamo invece la razza.

6.1 Sviluppi futuri

Questa tesi non pretende sicuramente di produrre risultati validi in ogni contesto, ma piuttosto si propone di fornire un fondamento da cui si possa sviluppare la propria idea sull'argomento trattato e approfondire nuove conoscenze.

Si spera infatti che indagini e ricerche future esplorino la tematica qui trattata, attraverso prospettive diverse e punti di vista complementari.

Le analisi condotte sul dataset Adult, applicando le definizioni che soddisfano il criterio di indipendenza, rappresentano soltanto un punto di partenza per affrontare le sfide connesse al concetto di fairness.

Le conclusioni emerse in questo lavoro potrebbero essere confermate

oppure contraddette da indagini sullo stesso dataset da altri punti di vista, come quello di altre definizioni. Tale ulteriore analisi, basata quindi sui criteri di separazione e sufficienza, sarebbe sicuramente diversa da quella appena condotta, pur mantenendo una significativa correlazione. Generalizzando, risulterebbe quindi interessante esaminare se lo stesso dataset può essere considerato fair o meno a seconda della definizione adottata o del criterio soddisfatto da tale definizione.

Inoltre, possiamo ipotizzare in modo abbastanza certo, che la proprietà di fairness non sia soddisfatta da molti altri dataset, a prescindere dalla prospettiva da cui la si analizzi. Quindi, così come è stato esaminato Adult, potrebbe essere interessante analizzare anche altri dataset, includendo sia quelli più famosi e comuni sia quelli più di specializzati e meno utilizzati.

Un obiettivo molto importante della ricerca in questo ambito, è quello di sviluppare soluzioni per migliorare il livello di fairness all'interno dei dati e per mitigare la discriminazione presente nei sistemi basati sull'apprendimento automatico [70]. Queste tecniche di miglioramento possono essere applicate ai dati prima o dopo l'esecuzione dell'algoritmo di machine learning: nel primo caso si parla di pre-processing, nel secondo di post-processing.

Un interessante studio potrebbe essere basato sull'analisi di Adult, o di altri dataset, dopo l'applicazione di tali interventi. Potrebbero essere confrontate varie tecniche di pre-processing, potenzialmente traendo conclusioni non banali e perfino inaspettate. Inoltre, sarebbe curioso e interessante analizzate le conseguenze delle varie tecniche di pre-processing sulle prestazioni dei modelli di apprendimento automatico.

6.2 Considerazioni finali

In conclusione, soprattutto alla luce dell'attuale sviluppo e del notevole utilizzo degli algoritmi di machine learning, si può giungere ad un'interessante osservazione. Quando utilizziamo Adult, o più in generale un

dataset non fair secondo un qualche attributo sensibile, per addestrare algoritmi di machine learning, si corre il rischio che il sistema risultante amplifichi i bias presenti al loro interno, conducendo a conclusioni non veritiere e quindi non affidabili.

Vista l'attualità e soprattutto l'enorme potenziale di questa tecnologia, è fondamentale che le aziende, gli sviluppatori, i ricercatori e i politici, collaborino in modo congiunto per affrontare questa sfida di primaria importanza. Trascurare gli aspetti legati al concetto di fairness potrebbe portare a limitare l'impiego dell'intelligenza artificiale in molti settori e campi, e quindi limitare il suo sviluppo e utilizzo. Affrontarli, d'altro canto, non porterebbe solo alla creazione di soluzioni migliori dal punto di vista tecnico, ma si estenderebbe alla costruzione di sistemi più affidabili, responsabili e rispettosi della dignità umana.

BIBLIOGRAFIA

[1] *About-UCI Machine Learning Repository*. UC Irvine. https://archive.ics.uci.edu/about. Data ultima visualizzazione: 2024-04-22.

[2] *European Parliament Adopts the AI Act: Implications for Culture*. https://cultureactioneurope.org/news/european-parliament-adopts-the-ai-act/. Data ultima visualizzazione: 2024-06-05.

[3] *HRExecutive*. https://www.hrexecutive.it. Data ultima visualizzazione: 2024-03-18.

[4] *Prometheus Blog*. https://www.prometheus-studio.it/prometheus_blog_wp/2019/09/29/cosa-e-una-matrice-di-confusione/. Data ultima visualizzazione: 2024-06-04.

[5] *proPublica*. https://www.propublica.org. Data ultima visualizzazione: 2024-04-06.

[6] *Python Pandas Tutorial: A Complete Introduction for Beginners*. learndatasci. https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/. Data ultima visualizzazione: 2024-04-20.

[7] *Turing Test*. https://www.andreaminini.net/computer-science/artificial-intelligence/turing-test. Data ultima visualizzazione: 2024-06-11.

[8] Dichiarazione universale dei diritti umani. 1948.

[9] *Reti neurali e vita artificiale*. Enciclopedia della Scienza e della Tecnica, 2007. https://www.treccani.it/enciclopedia/reti-neurali-e-vita-artificiale_(Enciclopedia-della-Scienza-e-della-Tecnica)/. Data ultima visualizzazione: 2024-03-17.

[10] Sariabili aleatorie. *Università degli studi di Siena*, 2011.

[11] *Variabile dipendente*. Treccani, 2012. https://www.treccani.it/enciclopedia/variabile-dipendente_(Dizionario-di-Economia-e-Finanza)/. Data ultima visualizzazione: 2024-03-25.

[12] *Variabile indipendente*. Treccani, 2012. https://www.treccani.it/enciclopedia/variabile-indipendente_%28Dizionario-di-Economia-e-Finanza%29/. Data ultima visualizzazione: 2024-03-25.

[13] *Unità statistica*. Treccani, 2013. https://www.treccani.it/enciclopedia/unita-statistica_(Enciclopedia-della-Matematica)/#. Data ultima visualizzazione: 2024-03-25.

[14] *Fairness*. Cambdridge Dictionary, 2014. https://dictionary.cambridge.org/it/dizionario/inglese-italiano/fairness. https://www.treccani.it/vocabolario/equita/. Data ultima visualizzazione: 2024-04-05.

[15] *Dove va l'intelligenza artificiale*. MIT Technology Review, 2015. https://www.linkiesta.it/2015/02/dove-va-lintelligenza-artificiale/. Data ultima visualizzazione: 2024-04-11.

[16] *La statistica negli studi clinici: distorsione (bias)*. Eupati, 2015. https://toolbox.eupati.eu/resources/la-statistica-negli-studi-clinici-distorsione-bias/?lang=it. Data ultima visualizzazione: 2024-04-16.

[17] *Regolamento (UE) 2016/679 del parlamento europeo e del consiglio del 27 aprile 2016*. Gazzetta ufficiale dell'Unione europea, 2016. https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016R0679. Data ultima visualizzazine: 2024-04-22.

[18] *Equità*. Treccani, 2018. https://www.treccani.it/vocabolario/equita/. Data ultima visualizzazione: 2024-04-03.

[19] *Imparziale*. Treccani, 2018. https://www.treccani.it/vocabolario/imparziale/. Data ultima visualizzazione: 2024-03-26.

[20] *L'intelligenza artificiale per l'Europa*. Commissione europea, 2018. https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A52018DC0237. Data ultima visualizzazione: 2024-03-21.

[21] *Onestà*. Treccani, 2018. https://www.treccani.it/vocabolario/onesta/. Data ultima visualizzazione: 2024-04-03.

[22] *Uguaglianza*. Treccani, 2018. https://www.treccani.it/enciclopedia/uguaglianza/. Data ultima visualizzazione: 2024-04-02.

[23] *Python Pandas Tutorial: A Complete Introduction for Beginners*. adult-all.csv, 2019. https://github.com/jbrownlee/Datasets/blob/master/adult-all.csv. Data di ultima visualizzazione: 2024-04-20.

[24] *Gruppo di esperti di alto livello sull'intelligenza artificiale*. Commissione europea, 2022. https://digital-strategy.ec.europa.eu/it/policies/expert-group-ai. Data ultima visualizzazione: 2024-04-19.

[25] *About pandas*. Pandas, 2024. https://pandas.pydata.org/about/. Data ultima visualizzazione: 2024-04-20.

[26] *Bias e intelligenza artificiale: tipi ed esempi*. Intelligenza Artificiale Italia, 2024. https://www.intelligenzaartificialeitalia.net/post/bias-e-intelligenza-artificiale-tipi-ed-esempi?wsdl. Data ultima visualizzazione: 2024-04-14.

[27] *The EU Artificial Intelligence Act*. EU Artificial Intelligence Act, 2024. https://artificialintelligenceact.eu. Data ultima visualizzazione: 2024-03-15.

[28] *History and License*. Python, 2024. https://docs.python.org/3/license.html. Data ultima visualizzazione: 2024-04-21.

[29] *Il Parlamento europeo approva la legge sull'intelligenza artificiale*. Parlamento europeo, 2024. https://www.europarl.europa.eu/news/it/press-room/20240308IPR19015/il-parlamento-europeo-approva-la-legge-sull-intelligenza-artificiale. Data ultima visualizzazione: 2024-03-12.

[30] *L'Intelligenza Artificiale nella selezione del personale*. Lavoro Diritti Europa, 2024. https://www.lavorodirittieuropa.it/dottrina/principi-e-fonti/1528-l-intelligenza-artificiale-nella-selezione-del-personale. Data ultima visualizzazione: 2024-04-16.

[31] *Neuralink, azienda miliardaria che punta al cervello*. Ansa, 2024. https://www.ansa.it/sito/notizie/cronaca/2024/01/30/neuralink-azienda-miliardaria-che-punta-al-cervello_81aaaa70-56cf-470a-91dd-8b4d0451837b.html. Data ultima visualizzazione: 2024-03-26.

[32] *NumPy: about us*. NumPy, 2024. https://numpy.org/about/. Data ultima visualizzazione: 2024-04-20.

[33] *PyPI*. Python Package Index, 2024. https://pypi.org. Data ultima visualizzazione: 2024-04-20.

[34] *shikit-learn*. shikit-learn, 2024. https://scikit-learn.org/stable/index.html. Data ultima visualizzazione: 2024-04-20.

[35] Nitin Agarwala and Rana Divyank Chaudhary. China's policy on science and technology: Implications for the next industrial transition. *India Quarterly*, 75(2):206–227, 2019.

[36] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[37] Yalin Başţanlar and Mustafa Özuysal. Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, pages 105–128, 2014.

[38] C Batini et al. Enciclopedia dei dati digitali, volume terzo: L'etica dei dati digitali: l'equità. 2022.

[39] Arie Ben-David and Eibe Frank. Accuracy of machine learning models versus "hand crafted" expert systems–a credit scoring case study. *Expert Systems with Applications*, 36(3):5264–5271, 2009.

[40] Ann Berrington, Peter Smith, and Patrick Sturgis. An overview of methods for the analysis of panel data. 2006.

[41] Omar Burzio. Machine learning per la qualità nell'industria 4.0, 2022.

[42] Piero Cappelletti. Medicina 4.0. un'introduzione. *La Rivista Italiana della Medicina di Laboratorio-Italian Journal of Laboratory Medicine*, 14(3):131–135, 2018.

[43] Carlo Casonato, Simone Penasa, et al. Intelligenza artificiale e medicina del domani. pages 553–586, 2021.

[44] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

[45] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[46] Alessandro Cimatti, Marco Pistore, Marco Roveri, and Paolo Traverso. Weak, strong, and strong cyclic planning via symbolic model checking. *Artificial Intelligence*, 147(1-2):35–84, 2003.

[47] Sandra Cimino. Intelligenza artificiale: implicazioni in termini di privacy, mercato e comportamento del consumatore, 2018.

[48] B Jack Copeland. *The essential turing*. Clarendon Press, 2004.

[49] I De Falco and D Maisto. Machine learning mediante programmazione genetica e induzione secondo. 2005.

[50] Tullio De Mauro. Grande dizionario italiano dell'uso. 2000.

[51] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[52] Allen Downey, Jeffrey Elkner, and Chris Meyers. *Pensare da informatico: imparare con Python*. Green Tea Press, 2003.

[53] Nexa Center for Internet and Society. Fairness e machine learning. il concetto di equità e relative formalizzazioni nel campo dell'apprendeimento automatico. 2018.

[54] Amir H Gandomi, Fang Chen, and Laith Abualigah. Machine learning technologies for big data analytics, 2022.

[55] Yihong Gong and Wei Xu. *Machine learning for multimedia content analysis*, volume 30. Springer Science & Business Media, 2007.

[56] Julian Gostoli. Sviluppo di un codice in python per l'analisi dell'efficacia di schermatura di materiali stratificati. 2021.

[57] Julian Gostoli. Sviluppo di un codice in python per l'analisi dell'efficacia di schermatura di materiali stratificati, 2022.

[58] Karli R Hochstatter, Wajiha Z Akhtar, Nabila El-Bassel, Ryan P Westergaard, and Marguerite E Burns. Racial disparities in use of non-emergency outpatient care by medicaid-eligible adults after release from prison: Wisconsin, 2015–2017. *Journal of substance abuse treatment*, 126:108484, 2021.

[59] James Johnson. Artificial intelligence and future warfare: implications for international security. *Defense & Security Analysis*, 35(2):147–169, 2019.

[60] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[61] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[62] Snezha SK Kazakova, Allison AD Dunne, Daan DB Bijwaard, Julien Gossé, Charles Hoffreumon, and Nicolas van Zeebroeck. European

enterprise survey on the use of technologies based on artificial intelligence. Technical report, ULB–Universite Libre de Bruxelles, 2020.

[63] Khuramova Farangiz Uchkun Kizi. Strong and weak artificial intelligence. 2022.

[64] Jakub Kufel, Katarzyna Bargieł-Łączek, Szymon Kocot, Maciej Koźlik, Wiktoria Bartnikowska, Michał Janik, Łukasz Czogalik, Piotr Dudek, Mikołaj Magiera, Anna Lis, et al. What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine. *Diagnostics*, 13(15):2582, 2023.

[65] Niklas Kühl, Max Schemmer, Marc Goutier, and Gerhard Satzger. Artificial intelligence and machine learning. *Electronic Markets*, 32(4):2235–2244, 2022.

[66] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

[67] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[68] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797, 2017.

[69] Chiara Macchiavello et al. Introduzione alle reti neurali. *atti del seminario tenuto all'università di Pavia il*, 17:20, 1992.

[70] Massimiliano Mancini. La fairness nel machine learning, 2023.

[71] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th*

ACM international conference on information & knowledge management, pages 2145–2148, 2020.

[72] John McCarthy et al. What is artificial intelligence. *Stanford University*, 2007.

[73] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[74] James Rufus McDonald, Stephen McArthur, Graeme Burt, and Jerry Zielinski. *Intelligent knowledge based systems in electrical power engineering*. Springer Science & Business Media, 1997.

[75] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[76] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[77] Tom M. Mitchell. *Machine learning*. McGraw-Hill, 1997.

[78] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[79] Francesco Pio Monaco. Bias negli algoritmi di machine learning-il caso degli algoritmi per l'assistenza sanitaria negli stati uniti, 2022.

[80] Nicholas Mro. Modelli di machine learning per la predizione di attacchi epilettici sulla base del segnale eeg, 2022.

[81] John Paul Mueller and Luca Massaron. *Machine learning for dummies*. John Wiley & Sons, 2021.

[82] Iqbal Muhammad and Zhu Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.

[83] Ravil I Mukhamediev, Yelena Popova, Yan Kuchin, Elena Zaitseva, Almas Kalimoldayev, Adilkhan Symagulov, Vitaly Levashenko, Farida Abdoldina, Viktors Gopejenko, Kirill Yakunin, et al. Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10(15):2552, 2022.

[84] N Musacchio, G Guaita, A Ozzello, MA Pellegrini, P Ponzani, R Zilich, and A De Micheli. Intelligenza artificiale e big data in ambito medico: prospettive, opportunità, criticità. *JAMD*, 21(3):1, 2018.

[85] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4(51-62):56, 2017.

[86] Nicolay Osalchuk. Raccolta ed analisi di dati relativi a pubblicazioni scientifiche e relative conferenze, 2022.

[87] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[88] Dana Pessach and Erez Shmueli. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Systems with Applications*, 185:115667, 2021.

[89] Paola Pucci, Giovanni Vecchio, et al. Mobilità e inclusione sociale. pianificare per vite sempre più mobili. In *Proceedings of XXI Conferenza Nazionale SIU. Confini, movimenti, luoghi. Politiche e progetti per città e territori in transizione"*, pages 107–113. Planum. The Journal of Urbanism, 2019.

[90] Michele Radicioni. *Sistema di visione stereoscopico per il packaging delle mele.* 2020.

[91] Filippo Rigotto. Visualizzazione dati con python: lo stato dell'arte, 2017.

[92] David E Rumelhart, Geoffrey E Hinton, James L McClelland, et al. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76):26, 1986.

[93] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[94] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE, 2013.

[95] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[96] Guido Sanguinetti. Machine learning: accuratezza, interpretabilità e incertezza. *Ithaca: Viaggio nella Scienza*, 2020(16):71–82, 2020.

[97] Jonathan Schaeffer. *One jump ahead: challenging human supremacy in checkers*. Springer Science & Business Media, 2013.

[98] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[99] Michaela Slussareff. O'neil, cathy. 2016. weapons of math destruction: How big data increases inequality and threatens democracy. crown., 2022.

[100] Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33):11629–11634, 2005.

[101] Santa Sozzi. Applicazione di un algoritmo non supervisionato basato su reti neurali per la rimozione del rumore speckle da immagini ecografiche. 2023.

[102] Oliver Stone. *Oliver Stone intervista Vladimir Putin*. Marsilio Editori spa, 2017.

[103] Alessandro Tatti. Metodi e tecniche per garantire la "fairness" e la "diversity" nell'estrazione di dati: un confronto, 2018.

[104] Vasil Teigens, Peter Skalfist, and Daniel Mikelsten. *Intelligenza artificiale: la quarta rivoluzione industriale*. Cambridge Stanford Books.

[105] KP Tripathi. A review on knowledge-based expert system: concept and architecture. *IJCA Special Issue on Artificial Intelligence Techniques-Novel Approaches & Practical Applications*, 4:19–23, 2011.

[106] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[107] Maria Cecilia Verri. Algoritmi e complessità. *Algoritmi e strutture dati*.

[108] David H Wolpert. Constructing a generalizer superior to nettalk via a mathematical theory of generalization. *Neural Networks*, 3(4):445–452, 1990.

[109] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V Vasilakos. Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237:350–361, 2017.