UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

TESI DI LAUREA

Anchoring bias: individuazione del bias all'interno dei processi decisionali

Relatore:
Alessio Malizia
D'Antoni

Candidato:
Giorgia

Correlatore:
Tommaso Turchi

ANNO ACCADEMICO 2021/2022

INDICE

ABSTRACT

I bias cognitivi non sono più oggetto di studio soltanto in ambiti umanistici, ma anche scientifici. I dati prodotti non sono altro che un riflesso del pensiero umano, riflettendo così anche la presenza di limitazioni nel ragionamento degli individui. Il bias di ancoraggio (o anchoring bias) rientra tra questi: il presente lavoro ha l'obiettivo di individuare il bias di ancoraggio riflesso all'interno dei processi decisionali, attraverso l'utilizzo di due dataset differenti e metodologie legate al mondo del machine learning. La tesi fa riferimento ad uno studio svolto da J. Echterhoff, M.Yarmand e J. McAuley all'Università di San Diego, ed ha lo scopo finale di superare il problema etico relativo all' utilizzo di algoritmi a scatola nera per l'individuazione del bias di ancoraggio all'interno dei processi decisionali: attraverso l'utilizzo dell'algoritmo di explainability LIME verrà spiegata la logica presente dietro la predizione dell'algoritmo utilizzato, permettendo così un possibile impiego degli algoritmi a scatola nera per l'automatizzazione di task aventi un impatto rilevante nella vita degli esseri umani.

INTRODUZIONE

Il seguente elaborato ha come obiettivo quello di fornire un possibile approccio nell'individuazione di bias all'interno dei dati, in particolare del bias di ancoraggio.
Al giorno d'oggi, infatti, il campo dell'intelligenza artificiale è utilizzato in qualsiasi ambito all'interno della nostra società: da quello finanziario a quello agricolo. L'innovazione tecnologica ha portato con sé un gran numero di vantaggi, permettendo l'automatizzazione dei processi attraverso l'utilizzo di algoritmi, analizzando un numero elevato di dati. Tuttavia, il progresso tecnologico non porta con sé soltanto vantaggi all'interno della società, ma anche svantaggi: oltre alla sostituzione dell'uomo con la macchina in alcuni settori lavorativi, vi è il problema del bias: l'algoritmo, infatti, si allena sui dati raccolti all'interno della nostra società, collezionati attraverso social media, smartphone, assistenti vocali ed altro, riflettendone e amplificandone tutti i pregiudizi presenti. Questo ha portato, nel corso della storia, alla presenza di discriminazioni da parte della macchina, come, ad esempio, quella di genere o di etnia, nei confronti di gruppi di individui, rendendo così necessaria, non solo un'opportuna regolamentazione, ma l'individuazione e la mitigazione di questi bias presenti anche all'interno di algoritmi a scatola nera e, quindi, spesso difficili da individuare prima che influiscano negativamente su diverse realtà sociali.
Una delle problematiche principali legate all'utilizzo dei dati per l'automatizzazione di sempre più numerose attività, infatti, è quella dell'utilizzo di algoritmi a scatola nera (o black box) che, se da una parte risultano avere una migliore performance nell'analizzare grandi quantità di dati, dall'altra non permettono di comprendere totalmente la logica presente dietro alle loro predizioni. Questo, dunque, può causare problemi etici riguardanti il loro utilizzo: automatizzare attività che hanno un impatto determinante nella vita delle persone, non conoscendo la logica dietro la predizione, può essere pericoloso poiché l'output risultante potrebbe contenere al suo interno bias di diverso tipo, discriminando una parte di individui. Per questo motivo, negli ultimi anni, attraverso la ricerca, sono stati sviluppati degli algoritmi di explainability in grado di sbirciare all'interno della scatola nera, osservando così la logica dietro la predizione ed eventuali bias presenti al suo interno.

L’elaborato è stato diviso in 4 capitoli principali, in modo da ripercorrere nel dettaglio la storia dei bias cognitivi e dell’intelligenza artificiale, fino ad arrivare a spiegarne la loro coesistenza: nel primo capitolo vengono osservati nel dettaglio i bias cognitivi, partendo dagli studi relativi alla loro origine e descrivendo alcuni dei bias cognitivi più frequenti nel ragionamento umano. Nel secondo capitolo, invece, ci si concentra sul ruolo che ricopre al giorno d’oggi l’intelligenza artificiale, partendo dalle sue origini. Il terzo capitolo crea un punto di unione tra i primi due, descrivendo le modalità con cui i bias cognitivi si riflettono all’interno della macchina e le relative conseguenze. Infine, il quarto ed ultimo capitolo, si concentra sulla possibile individuazione del bias di ancoraggio all’interno di due dataset differenti, per poi utilizzare un algoritmo di explainability al fine di spiegare la predizione di ancoraggio.

1. I Bias Cognitivi

La maggior parte delle teorie classiche riguardanti il ragionamento umano propone una visione secondo la quale gli esseri umani tendono a ricercare sempre la risposta ottimale nel momento in cui prendono decisioni e formano giudizi. Secondo questo punto di vista, prevalente in molte scienze cognitive, le persone si comportano come agenti razionali in grado di risolvere problemi cognitivi semplici e complessi e di massimizzare i benefici che possono ottenere dalle loro interazioni con l'ambiente. In generale, si considererebbero i potenziali costi e vantaggi delle proprie attività prima di selezionare la scelta complessivamente più vantaggiosa. Ciò comporta la considerazione di tutte le informazioni necessarie per la risoluzione del problema, escludendo qualsiasi informazione irrilevante che possa contaminare la scelta. Questa idea di razionalità ha sostenuto interi campi di studio nelle scienze sociali.
Questo approccio tradizionale, tuttavia, è stato messo in discussione negli ultimi decenni: una quantità crescente di prove empiriche rivela che i giudizi e le azioni delle persone sono spesso tutt'altro che razionali essendo in realtà influenzati da circostanze apparentemente insignificanti[1]. Inoltre, queste deviazioni dal pensiero logico sono tipicamente sistematiche: le persone falliscono ripetutamente nello stesso tipo di situazione, commettendo lo stesso errore. In altre parole, gli individui sembrano essere illogici in modi prevedibili. Di conseguenza, una teoria che mira a replicare il giudizio umano e il processo decisionale deve essere in grado di spiegare questi casi di pregiudizi persistenti o bias cognitivi che possono essere considerati come una divergenza sistematica nel processo decisionale.

1.1 Etimologia e storia

Le origini del termine *bias* risalgono al 1560: il termine veniva utilizzato nel vecchio gioco delle bocce e faceva riferimento a palline realizzate con un peso maggiore su un lato, che le faceva curvare obliquamente; da qui l'uso figurativo per indicare una tendenza unilaterale della mente.

[1] M. Gladwell, *Blink: The power of thinking without thinking*, 2006.

Il termine *bias cognitivo*, invece, trova le sue origini nel 1970, coniato per la prima volta da Amos Tversky e Daniel Kahneman, psicologi israeliani che hanno usato queste parole per descrivere diversi modelli di pensiero delle persone che risultano imperfetti in risposta a problemi che prevedono un giudizio o una decisione.

Viene naturale chiedersi per quale motivo questi modelli abbiano in primo luogo avuto origine: la teoria della gestione degli errori[2] di Haselton e Nettle definisce la formazione dei bias cognitivi come una necessità evolutiva, in grado di offrire vantaggi per la sopravvivenza: situazioni di pressione in cui è necessario prendere in fretta decisioni di vita o di morte sono riconducibili già ai tempi dei nostri antenati, quando, ad esempio, si trovavano di fronte ad un possibile pericolo. Queste condizioni non solo favoriscono lo sviluppo di veloci meccanismi decisionali ma producono anche il cosiddetto errore meno costoso: una decisione non ottimale ma con un errore basso nel risultato. Molti bias cognitivi sembrano favorire sistematicamente la conclusione che si allinea con l'errore meno costoso. Inoltre, l'esibizione di determinati bias cognitivi può produrre anche altri tipi di benefici, in particolare in termini emotivi come del caso dell'illusione del controllo[3]: l'errata convinzione che si possa esercitare un controllo su risultati che sono in realtà incontrollabili. Le conseguenze emotive sono degne di nota: una persona che pensa non ci sia nulla che si possa fare per influenzare un risultato rilevante potrebbe sentirsi disperata, essendo una triste consapevolezza non avere il controllo sulla propria vita. Al contrario, coloro che sviluppano l'illusione del controllo si attribuiscono qualsiasi risultato positivo che possa accadere, quindi si sentono fiduciosi e al sicuro. Inoltre, si sentiranno anche motivati a continuare a provare e produrre azioni per influenzare l'ambiente. In sintesi, alcuni bias cognitivi sembrano essere associati a esiti positivi, o alla minimizzazione a lungo termine di errori costosi, rappresentando un vantaggio evolutivo.

Kahneman e Tversky[4] sono stati tra i primi a dimostrare l'esistenza di errori di giudizio, o "illusioni cognitive", all'interno dei ragionamenti umani, così come anche

---

[2] M. G. Haselton, D. Nettle, *The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases*, 2006.

[3] E. J. Langer, *The illusion of control*, 1975.

[4] D. Kahneman, A. Tversky, *Subjective probability: A judgment of representativeness*, 1972.

Peter Wason. Tra gli esperimenti più famosi troviamo: il Linda problem, l'hospital problem e il Wason selection task.

Il Linda problem (o fallacia della congiunzione), ideato da Kahneman e Tversky, si basa sulla probabilità congiunta[5] : Linda è una donna di 31 anni, single e brillante. Laureata in filosofia, durante il suo periodo di studentessa, si è anche impegnata in problemi di discriminazione e giustizia sociale partecipando a diverse manifestazioni, Viene chiesto poi ai partecipanti quale tra due diverse affermazioni sia la più probabile: a) Linda lavora in banca, oppure b) Linda lavora in banca ed è attiva nel movimento femminista. La probabilità che due eventi si verifichino allo stesso momento si calcola moltiplicando tra loro le probabilità dei singoli eventi, operazione che potrebbe far pensare di ottenere dunque un numero più elevato scegliendo l'opzione b. Tuttavia, il prodotto di due numeri compresi tra 0 e 1 (la probabilità assume un valore nell'intervallo tra 0 e 1) risulta sempre in un numero minore di ciascuno dei due fattori, motivo per il quale la prima affermazione risulta essere quella corretta.

Nella loro ricerca gli autori indicano che circa l'85% dei votanti sceglie erroneamente la seconda opzione. Una possibile spiegazione del fenomeno risiede nel modo in cui il personaggio di Linda viene presentato nell'esperimento: Linda viene inizialmente rappresentata come una persona attiva nelle questioni sociali, portando le persone a pensarla come una persona tuttora coinvolta nel sociale, oltre che nel ruolo lavorativo di banchiera, a tal punto da violare le leggi della probabilità.

Gli autori suggeriscono che una possibile riduzione del fenomeno della fallacia della congiunzione, osservandone le cause, potrebbe essere o quella di cambiare la modalità in cui viene descritta Linda e posto il problema, oppure quella di cambiare il format delle opzioni di risposta sostituendo le affermazioni con una stima delle probabilità delle due opzioni. Inoltre, invitare a immaginare un numero elevato di donne conforme alla descrizione di Linda, farebbe realizzare ai partecipanti che è molto più probabile che ci siano donne che lavorano in banca, che donne che lavorano in banca e sono anche attive nel movimento femminista.

[5] Probabilità che più eventi si verifichino contemporaneamente.

Il Wason selection task (o compito di selezione di Wason), ideato da Peter Wason nel 1966, è un problema logico composto da quattro carte ed una regola. Le carte mostrano su una facciata i simboli *A*, *K*, *4* e *7*, la regola prevede che se una carta ha su una delle due facciate una vocale, allora sull'altra facciata sarà presente un numero pari. L'obiettivo posto ai partecipanti è quello di verificare la veridicità dell'enunciato scegliendo quali carte girare tra le quattro, scegliendone massimo due. Al fine di constatare se la regola sia vera o meno, le carte che andrebbero girate sono quelle aventi sulla facciata i simboli *A* e *7*, poiché sono le uniche carte che potrebbero potenzialmente violare la regola. L'interesse sviluppatosi per questo tipo di problema deriva dal fatto che malapena il 10% dei partecipanti ha saputo fornire una risposta corretta: il 46% ha scelto di girare le carte aventi sulla facciata i simboli *A* e *4*, mentre il 33% ha scelto di girare soltanto la carta con il simbolo *A.* Sembra essere chiaro a tutti i partecipanti che la carta con il simbolo *A* vada girata, dal momento che se ci fosse un numero dispari sull'altra facciata la regola verrebbe violata. Tuttavia, girare la carta con il simbolo *4* non è corretto né necessario dal momento che, anche se ci fosse una consonante dall'altro lato: non dicendo nulla riguardo al retro delle consonanti la regola non verrebbe violata. Appare così cruciale girare la carta con il simbolo *7* poiché se ci fosse una vocale sull'altro lato verrebbe smentita la regola.

Quando l'autore cerca di convincere i partecipanti dei loro errori, incontra una certa resistenza, poiché essi continuano ad affermare che girare la carta con il simbolo *7* non sia necessario. Una spiegazione cognitiva a questo errore è che le persone tendono solitamente a confermare le proprie affermazioni attraverso l'utilizzo di nuove informazioni, piuttosto che provare a smentirle: chiunque giri la carta con il simbolo *A* ha la possibilità di confermare la regola *"*se vocale, allora numero pari", mentre chiunque giri la carta con il simbolo *7* può al massimo smentirla. Questo tipo di processo cognitivo sembra essere anche alla base della formazione delle teorie complottiste[6].

Le possibilità di risolvere il problema possono aumentare significatamene se si pongono esempi con problemi reali: ad esempio, utilizzare l'implicazione "Se sto davanti la Torre Eiffel, allora sono a Parigi" e il suo corrispettivo inverso "Se non sono

[6] Y. Majima, *Belief in pseudoscience, cognitive style and science literacy,* 2015.

a Parigi, non posso stare davanti alla Torre Eiffel" permette di cogliere più facilmente la logica che si nasconde dietro la regola e a risolvere correttamente compiti simili

L'hospital problem, ideato come il Linda problem da Kahneman e Tversky nel 1972, è un problema matematico basato sulla legge dei grandi numeri[7]: una città è fornita di due ospedali. Nel primo ospedale, il più grande, 45 bambini nascono ogni giorno. Nell'ospedale più piccolo, invece, nascono circa 15 bambini al giorno. Di tutti questi bambini solitamente circa il 50% sono maschi. Tuttavia, la percentuale varia di giorno in giorno: a volte può essere poco più del 50%, a volte poco meno. Viene poi sottolineato che per la durata di un anno entrambi gli ospedali hanno registrato i giorni in cui più del 60% delle nascite erano di sesso maschile.
Con questi dati a loro disposizione i partecipanti devono indovinare quale ospedale, secondo loro, ha registrato il maggior numero di giorni nell'arco dell'anno in cui la percentuale di bambini maschi era superiore al 60%. I votanti possono o scegliere tra uno dei due ospedali, oppure scegliere una terza opzione: entrambi gli ospedali (con uno scarto massimo del 5% l'uno dall'altro).
Il 56% sceglie la terza opzione, mentre il 22% sceglie una tra le prime due. Tuttavia, secondo la legge dei grandi numeri, è più probabile che l'ospedale più grande presenti un rapporto tra i due sessi vicino al 50%, motivo per il quale l'opzione corretta risiede nell'ospedale più piccolo. Secondo Tversky and Kahneman la causa per cui la maggior parte dei partecipanti scelgono la terza opzione sta nell'euristica della rappresentatività: siccome la somiglianza non dipende dalla dimensione del campione, le persone che seguono questa euristica ignorano l'importanza della dimensione del campione. Parlando in termini statistici, i partecipanti giudicano erroneamente il campione (in questo caso l'ospedale più piccolo) come avente le stesse proprietà della popolazione da cui proviene (l'ospedale più grande), gli eventi riguardanti i due ospedali, infatti, sono descritti con la stessa statistica per quanto riguarda le nascite e dunque considerati equamente rappresentativi della popolazione generale.

[7] La legge dei grandi numeri prevede che se un test viene eseguito ripetutamente (come, ad esempio il lancio, di un dado), allora la frequenza che un evento si presenti (es. il fatto che esce il numero 3) tenderà ad una costante, ovvero la probabilità che quel determinato evento si verifichi.

Anche nella ricerca[8] di Howard Wainer e Harris L. Zwerling viene affrontato un tema simile: gli autori dimostrano che le percentuali di cancro al rene sono più basse nelle contee che sono per lo più rurali, scarsamente popolate e situate in stati tradizionalmente repubblicani nel Midwest, nel sud e nell'ovest. Allo stesso tempo, ad avere le percentuali più alte, sono contee con le stesse identiche caratteristiche. Sebbene il fenomeno potrebbe essere attribuito a vari fattori ambientali ed economici, gli autori sostengono che è dovuto alla dimensione del campione: a causa della piccola dimensione del campione, è più probabile che l'incidenza di un certo tipo di cancro nelle piccole contee rurali sia più lontana dalla media, in una direzione o nell'altra, rispetto all'incidenza dello stesso tipo di cancro in aree urbane molto più densamente popolate.

Le ricerche riguardanti i bias cognitivi hanno attirato l'attenzione di molti ricercatori nelle diverse discipline e hanno anche influenzato notevolmente il campo della medicina, della giurisprudenza e dell'economia, rendendo evidente come anche gli esperti, nei loro rispettivi campi, possono incorrere in errori logici e statistici.
Negli anni Novanta è stato avviato un contro movimento avviato principalmente dallo psicologo tedesco Gerd Gigerenzer: nella sua ricerca *enlightening program* (programma illuminante), sono stati sviluppati strumenti cognitivi che hanno l'obiettivo di consentire alle persone di comprendere le diverse illusioni cognitive e i rompicapi statistici. L'idea alla base di questa ricerca è semplicemente quella di cambiare la rappresentazione delle informazioni fornite, e non di addestrare le persone alla risoluzione dei problemi ancor prima di presentare un problema.

1.2 Principali meccanismi e contesti di sviluppo

Nel corso degli anni, la ricerca si è concentrata maggiormente sullo sviluppo di una tassonomia per i bias individuati sperimentalmente. Tuttavia, diversi autori hanno fornito quadri concettuali coerenti per comprendere cosa hanno in comune tutti questi bias e come si manifestano.

---

[8] H. Wainer, H. L. Zwerling**,** *Evidence That Smaller Schools Do Not Improve Student Achievement*, 2006.

In primo luogo, come accennato anche precedentemente, la limitata capacità di elaborazione della mente umana è una chiara spiegazione a molti pregiudizi, o bias, cognitivi documentati: poiché i ricordi delle persone non hanno una capacità illimitata, non è possibile esaminare una quantità molto grande di informazioni mentre si effettua un'inferenza o una decisione, nonostante ogni cosa abbia la sua importanza. Piuttosto, si è costretti a concentrarsi su un sottoinsieme delle informazioni disponibili, che spesso non si è nemmeno in grado di comprendere appieno. Di conseguenza, nei casi più complicati, la risposta ideale e ottimale da un punto di vista razionale, è fuori portata e si può aspirare soltanto ad una razionalità limitata, detta anche *bounded rationality*[9], prendendo la decisione migliore con soltanto una quantità ristretta di informazioni.

L'emozione (o l'affetto) è un'altra possibile fonte di sviluppo per i bias cognitivi. La ricerca riguardante il processo decisionale ha definito la razionalità come una "coerenza formale", conforme ai principi della teoria della probabilità e dell'utilità. Le emozioni vengono quindi escluse da tale definizione poiché considerate un elemento in grado di contaminare i risultati di un processo decisionale alterandolo. Tuttavia, ulteriori ricerche (come quella degli autori Bechera e Damasio in *The somatic marker hypothesis*[10]*)* rivelano che le emozioni svolgono un ruolo significativo nel processo decisionale e che le decisioni non sarebbero mai del tutto ottimali senza valutazioni emotive. In fondo, le emozioni sono significative perché influenzano il comportamento e, tale influenza, permette di spiegare una varietà di bias cognitivi come, ad esempio, quello di avversione alla perdita: definito come la preferenza ad evitare perdite piuttosto che raccogliere guadagni di pari valore. Questo comportamento può essere motivato dal fatto che le emozioni negative tendono ad avere un impatto emotivo maggiore rispetto a quelle più piacevoli. In altri casi l'emozione influenza il giudizio morale come, ad esempio, nel noto *dilemma del*

[9] Il termine fu coniato per la prima volta da Herbert A. Simon al fine di indicare le limitazioni alle quali è soggetta la razionalità umana durante un processo decisionale, come ad esempio il tempo a disposizione, la quantità di informazioni conosciuta, ecc.

[10] Pubblicato nel 2005, l'articolo si pone come obiettivo quello di dimostrare il ruolo chiave delle emozioni nel processo decisionale in ambito economico.

carrello[11]: un carrello ha perso il controllo e sta precipitando giù per le rotaie. Cinque individui sono legati alla ferrovia rischiando la propria vita: il partecipante può scegliere di tirare una leva per dirigere il carrello verso un binario laterale dove invece è intrappolato soltanto un passeggero. La maggior parte degli individui sceglie di tirare la leva al fine di ottenere il minor numero di perdite possibile. Tuttavia, le decisioni dei partecipanti in queste situazioni sono vulnerabili alle manipolazioni affettive: ad esempio, se la persona sul binario laterale è il parente più stretto di un partecipante o il suo partner, ciò potrebbe influenzare la scelta. Di conseguenza, le emozioni possono causare alcune deviazioni sistematiche dalla norma razionale.

Anche i segnali sociali sono un elemento in grado di alterare il ragionamento formando alcuni bias cognitivi: il bandwagon bias, ad esempio, è caratterizzato dal desiderio delle persone di aderire a convinzioni precedentemente espresse da altri, risultando in un grande impatto sulle azioni collettive come il voto durante le elezioni.

Infine, il programma di ricerca sull'euristica di Kahneman e Tversky è forse il tentativo più efficace di stabilire un quadro coeso per la comprensione dei bias cognitivi. La logica di questo approccio prevede che fare scelte razionali non è sempre possibile, o addirittura auspicabile, per diversi motivi: ci vuole tempo per raccogliere e ponderare con impegno tutte le prove per risolvere un problema e spesso, in un problema, un'approssimazione della soluzione ottimale può essere considerata discreta, che continuare a lavorare per ottenere la soluzione ottimale sarebbe costoso. Di conseguenza, la mente impiega euristiche, o scorciatoie mentali, per raggiungere un giudizio in modo rapido ed economico. Un'euristica è una regola di base che genera una risposta veloce con il minimo sforzo piuttosto che catturare il problema in tutta la sua complessità o arrivare alla soluzione ideale. Nella ricerca pubblicata nel 1982[12] gli stessi autori hanno identificato varie euristiche che possono essere alla base di molti bias cognitivi: l'euristica della rappresentatività, l'euristica della disponibilità e l'euristica dell'ancoraggio.

---

[11] A. Blesche-Rechek, *Evolution and the trolley problem: people save five over one unless the one is young, genetically related, or a romantic partner,* 2010.

[12] D. Kahneman, A. Tversky, *Judgment under uncertainty: Heuristics and biases,* 1982.

L’euristica della rappresentatività, già nominata precedentemente, si basa sulla somiglianza o appartenenza: gli autori sottolineano che nel rispondere a domande probabilistiche come “Quale è la probabilità che l'oggetto A appartenga alla classe B?” o “Qual è la probabilità che l'evento A abbia origine dall’evento B?” le persone in genere fanno affidamento all'euristica della rappresentatività, in cui le probabilità sono stimate in base al grado di somiglianza tra l’evento A e l’evento B: solitamente se i due eventi si somigliano le probabilità che l’evento A abbia origine da B viene considerata molto alta.

Inoltre, uno dei fattori chiave nel calcolo delle probabilità è la *probabilità a priori* [13] di un evento, non presa in considerazione quando la probabilità viene stimata basandosi sulla somiglianza.

In sintesi, secondo l’euristica della rappresentatività, quando un esemplare è percepito come rappresentativo del gruppo, all'esemplare vengono attribuite tutte le caratteristiche tipiche del gruppo.

L'euristica della disponibilità è determinata dalla velocità con cui una rappresentazione arriva alla mente. Se un concetto è semplice da evocare o concepire, viene erroneamente valutato come più probabile che si verifichi: ad esempio, si potrebbe valutare il rischio di infarto tra persone di mezza età in base al ricordo di tali ricorrenze tra i propri conoscenti. La disponibilità può essere però considerata anche come un indizio utile per valutare la frequenza o probabilità di un evento dal momento che le istanze di classi molto grandi vengono solitamente richiamate più velocemente delle istanze di classi meno frequenti, nonostante sia influenzata da fattori diversi dalla frequenza e dalla probabilità.

L’euristica dell’ancoraggio è talvolta vista come un sottoinsieme dell'euristica della disponibilità.

In molte situazioni, le persone fanno stime partendo da un valore iniziale che viene aggiustato per fornire la risposta finale. Il valore iniziale, o punto di partenza, può essere suggerito dalla formulazione del problema, oppure può essere il risultato di un calcolo parziale. In entrambi i casi, gli aggiustamenti sono generalmente insufficienti: diversi punti di partenza producono stime diverse, che sono distorte verso i valori

[13] La probabilità a priori è la probabilità che un evento si verifichi prima che vengano raccolti nuovi dati.

iniziali. In una dimostrazione dell'effetto, riportata nella ricerca di Kahneman e Tversky, ai soggetti viene chiesto di stimare diverse quantità (ad esempio il numero di popolazioni che non possiedono l'elettricità) partendo da un numero compreso tra 0 e 100 ottenuto facendo girare una ruota della fortuna in presenza dei soggetti. I soggetti sono poi incaricati di indicare se il numero ottenuto è secondo loro superiore o inferiore al valore reale della quantità, di cui devono poi stimarne il valore spostandosi verso l'alto o verso il basso dal numero precedentemente dato dalla ruota. Nell'esperimento il numero di partenza ottenuto girando la ruota è diverso per ciascun gruppo ed ha un effetto marcato sulle stime. Ad esempio, le stime mediane della percentuale di paesi africani nelle Nazioni Unite erano 25 e 45 per i gruppi che ricevevano rispettivamente 10 e 65 come punti di partenza. La probabilità di un evento elementare dichiarata inizialmente fornisce un punto di partenza naturale per la stima delle probabilità di eventi sia congiuntivi che disgiuntivi, poiché l'aggiustamento dal punto di partenza è tipicamente insufficiente, le stime finali restano solitamente troppo vicine alle probabilità degli eventi elementari in entrambi i casi. Il tema dell'ancoraggio è un elemento chiave per la comprensione dell'intero elaborato e verrà per questo approfondito anche nei capitoli a seguire.

1.3 Principali bias cognitivi

I bias cognitivi sono dunque euristiche inefficaci, pregiudizi astratti che non si generano su dati reali.
Di seguito, verranno elencati e descritti alcuni tra i bias cognitivi esistenti, dal momento che i bias riconosciuti sono al momento più di 200.

- **Bias di ancoraggio (o anchoring bias):** già osservato precedentemente e oggetto di analisi in questa tesi, il bias di ancoraggio consiste in errori di percezione nelle scelte decisionali e sono spesso studiati ed utilizzati dagli esperti di marketing nelle vendite. Questa distorsione nella scelta trova le sue radici, infatti, nel punto di ancoraggio o punto di partenza, che può essere ad esempio il prezzo di un determinato prodotto: ogni prezzo osservato successivamente verrà paragonato al primo e la scelta finale sarà quindi calibrata partendo da esso. Questo meccanismo spiega anche, ed esempio, il motivo per cui nei motori di ricerca si paga per essere al primo posto nelle

ricerche, diventando il punto di ancoraggio dell'utente. Un altro esempio facilmente comprensibile è la presenza del prezzo originario nel momento in cui sun prodotto viene scontato: anziché presentare soltanto il prodotto con il suo nuovo prezzo, si preferisce mettere in evidenza il prezzo di partenza mantenendolo ben visibile al consumatore in modo da svolgere la funzione di ancoraggio. Ovviamente, maggiore sarà la differenza tra il prezzo di origine e quello scontato, più si avrà la percezione che l'affare sia buono aumentando così le probabilità di vendita. Allo stesso modo anche quando si procede all'acquisto di un oggetto costoso online, come ad esempio un laptop, si crea un punto di ancoraggio rappresentato dal prezzo del prodotto presente nel carrello: durante la finalizzazione della procedura d'acquisto, avendo sempre davanti il prezzo del prodotto iniziale, vengono solitamente suggeriti prodotti accessorio di prezzo inferiore acquistati da altri utenti insieme al primo prodotto: la grande differenza di prezzo farà sembrare il prezzo dell'accessorio un buon affare. Attraverso l'ancoraggio, la decisione finale si avvicina sempre al primo dato ancorato.

- **Bias di costo:** esistono bias di costo di vario tipo, tra i più rappresentativi si piò trovare l'effetto dotazione: consiste nel dare un valore differente e maggiore ad un oggetto in proprio possesso rispetto ad un oggetto che non lo è. Il meccanismo viene applicato anche se si prende in considerazione lo stesso identico oggetto. Come già visto precedentemente l'affetto gioca un ruolo fondamentale: è dunque l'avversione alla perdita a dare origine all'effetto dotazione. Un esempio per comprendere al meglio la fallacia nel ragionamento in questo tipo di bias può essere osservato prendendo come riferimento uno spettacolo teatrale: se lo spettacolo per la quale si è pagato il biglietto non è di gradimento, chi ha effettuato l'acquisto sarà più motivato a rimanere dal momento che ha speso dei soldi per vederlo. Se nella stessa situazione il biglietto fosse stato ottenuto gratuitamente, probabilmente lo spettatore se ne sarebbe andato.

- **Bias di desiderio:** trai bias di desiderio più famosi ci sono l'effetto della mano calda, l'effetto alone e l'effetto della mera esposizione.

L’effetto della mano calda ha origine dalla convinzione che vi sia una reale influenza degli eventi passati su un evento futuro generato dal caso: questo meccanismo porta a credere a chi ha avuto fortuna in una determinata situazione, di poterla allora avere anche successivamente. Un esempio attuale dell’effetto della mano, e fonte di patologie come la ludopatia, è il gioco d’azzardo. Il vizio del gioco è diventato negli ultimi anni oggetto di preoccupazione in diversi paesi: a spingere le persone a giocare continuamente fino ad avere grandi perdite è la convinzione di poter vincere sempre di più dopo aver casualmente avuto successo nelle prime puntate.

L’effetto Alone consiste nella valutazione positiva o negativa di oggetti o persone basandosi soltanto su alcune loro caratteristiche, ritenendole influenti in giudizi che non hanno alcun reale collegamento logico con queste. L’effetto potrebbe essere ricollegato alla famosa espressione “l’abito non fa il monaco”[14]: si tende, ad esempio, a giudicare un uomo come più intelligente di un altro soltanto perché vestito in modo migliore, oppure si tende a giudicare una donna meno intelligente soltanto perché vestita in abiti più succinti.

L’effetto della mera esposizione, infine, è un fenomeno psicologico per cui si tende ad avere una preferenza verso cose o persone soltanto perché ci si è stati più spesso a contatto: più si ascolta o vede qualcosa, maggiori sono le possibilità che questa cosa inizi a piacere. Questo accade perché nel tempo viene sviluppata una sorta di familiarità, sensazione positiva per l’essere umano, che tende ad evitare sensazioni non gradite come ansia e paura.

- **Bias di Framing:** I Bias di Framing si verificano quando il cervello esprime giudizi sulle informazioni in base a come queste vengono presentate. Nel marketing, l'effetto framing viene spesso utilizzato per influenzare i clienti e i loro acquisti: il bias sfrutta la propensione delle persone a rispondere in modo diverso davanti alle stesse informazioni, a seconda dal loro inserimento in un contesto positivo o negativo. Esistono molteplici bias appartenenti a questa categoria, verranno analizzati di seguito alcuni tra i più frequenti:

[14] Proverbio italiano che invita a non giudicare una persona dalle apparenze, evitando giudizi superficiali che non sempre corrispondono alla realtà.

– Il bias di conferma, già osservato nell'esperimento di Wason, è la tendenza ad elaborare affermazioni cercando soltanto informazioni coerenti con le proprie opinioni. Questo approccio parziale al processo decisionale è spesso non intenzionale e si traduce nel rifiuto verso fatti che contraddicono le proprie idee: si tende infatti ad evitare individui, gruppi o situazioni sociali che contraddicono il proprio modo di pensare, perché ciò produrrebbe pensieri tra loro contrastanti e, di conseguenza, creare una situazione di disagio. Quando un problema è altamente essenziale o rilevante per sé stessi, le persone sono più inclini ad analizzare informazioni che supportino le proprie opinioni. Dal momento che tutti tendono a ricercare contesti e informazioni in linea con sé stessi, il bias di conferma è uno dei bias maggiormente studiati nel campo della psicologia cognitiva. Questo meccanismo può essere visto come una forma di autoinganno, causata da un'esigenza di difesa personale: le idee e i valori sono infatti alla base della propria identità personale. Un'altra possibile causa si può trovare nel senso di appartenenza: le idee in cui si crede sono spesso le stesse del gruppo a cui si appartiene e, se contraddette, il senso di appartenenza ad esso viene a mancare. Il luogo dove maggiormente si sviluppa il bias di conferma è online: attraverso l'utilizzo di algoritmi di intelligenza artificiale, allenati attraverso dati forniti dagli utenti, è possibile creare un profiling[15] per ogni persona che utilizza una determinata piattaforma. La creazione di un profiling permette di mostrare all'utente un numero sempre maggiore di contenuti in linea con le sue idee. Per quanto l'idea di ricevere così facilmente contenuti di proprio interesse possa sembrare piacevole, può anche essere molto pericoloso: se, ad esempio, una persona si convince attraverso articoli complottisti di un'idea sbagliata o non del tutto vera, l'algoritmo continuerà a

[15] Registrazione e l'analisi delle caratteristiche psicologiche e comportamentali di una persona, in modo da identificare categorie di persone e prevederne le attività.

fornire informazioni che la avvalorano, rendendo sempre più difficile per l'utente ottenere un'informazione corretta.

- L'errore di attribuzione, invece, è la tendenza ad attribuire la causa del comportamento di una persona alla sua personalità piuttosto che all'ambiente e alle situazioni a lei circostanti. Questo ragionamento è di base errato poiché il contesto sociale, in ogni situazione, svolge un ruolo fondamentale nella comprensione dell'atteggiamento e delle scelte di una persona, facendo parte di essa dal momento in cui viene al mondo. Se nel giudicare il comportamento di qualcuno in una determinata situazione si scinde dal contesto in cui si trova, si ottiene una visione soltanto parziale del quadro d'insieme. Questo errore svolge un ruolo fondamentale nel modo in cui gli esseri umani si percepiscono e interagiscono tra loro. Ovviamente la situazione cambia a seconda del punto di vista: quando si è attori all'interno di una situazione, il contesto sembrerà essere fondamentale per spiegare le proprie scelte, al contrario, quando si è spettatori, sarà molto più semplice non tenerlo in considerazione. In determinati casi però, come ad esempio quando si vuole mantenere alta la propria autostima, l'attore tende ad attribuire la causa dei suoi successi alla sua personalità o impegno, e non al contesto intorno a sé, in caso di insuccesso, invece, l'attore attribuisce la colpa agli eventi circostanti. L'esempio tipico è quello di un uomo che assiste ad un incidente automobilistico: dal suo punto di vista le cause dell'incidente sono molteplici e tutte attribuibili a diversi tratti della personalità del conducente come, ad esempio, l'egoismo nel voler superare o l'incapacità nel guidare. In una situazione in cui invece è lui a causare l'incidente, la causa verrà ricercata in fattori esterni come, ad esempio, una strada dissestata o un cartello non visibile.

- L'In-group Bias (o favoritismo di gruppo) è la tendenza che hanno gli esseri umani ad essere solitamente più disponibili e positivi nei confronti dei membri del proprio gruppo rispetto a membri di un gruppo esterno. L'In-group bias trova le sue origini nella primitiva necessità

dell’uomo avere un’identità di gruppo. L’Etnia, i partiti politici e le istituzioni religiose sono tutti esempi attuali di identità di gruppo nella realtà sociale. In genere, anche all’interno di contesti scolastici o lavorativi, quando si viene divisi casualmente in gruppi è più probabile che si favoriscano i membri appartenenti al proprio gruppo, piuttosto che i membri appartenenti ad un altro. Il pregiudizio che si forma all’interno del gruppo, tuttavia, non è costante ma subordinato al cambiamento di idee e necessità dei suoi membri. Le elezioni ne sono un buon esempio: i membri di un partito politico si dividono in fazioni contrastanti tra loro che sostengono diversi candidati dello stesso partito. Tuttavia, una volta che il rappresentante di una fazione viene scelto per candidarsi, i membri del partito si fondono in un gruppo unico che ha l’obiettivo di sostenere il candidato scelto, spostando il loro pregiudizio verso il candidato del partito avversario.

L’In-group bias può avere gravi conseguenze nel mondo reale, in particolare per le persone appartenenti a gruppi minoritari: le persone con un alto livello di razzismo sono solitamente molto veloci nel giudicare un errore commesso da una persona afroamericana, così quanto lo sono nel giustificare invece un cattivo comportamento adottato da un cittadino americano o europeo. Questo accade anche perché l’In-group bias può portare ad essere più indulgenti nei confronti dei membri del proprio gruppo quando fanno qualcosa di sbagliato. Molte sono anche le implicazioni che si riflettono nel comportamento morale: le persone sono più disposte a mentire o imbrogliare per avvantaggiare il proprio gruppo anche quando non hanno intenzione di guadagnare nulla da questa disonestà. Questo favoritismo può portare a prendere decisioni sbagliate, soprattutto se il comportamento ha origine da una mancanza di autostima nell’individuo.

Esistono diverse teorie sul motivo sull’origine di questi pregiudizi all'interno del gruppo, ma una delle più importanti è quella dell'identità sociale: la concezione di identità si basa in parte sulle categorie sociali a cui si appartiene. Non tutte le categorie sono ugualmente importanti, ma tutte contribuiscono all'idea che si ha di sé stessi e del ruolo che si svolge all’interno della società.

- Il Bias della correlazione illusoria è la percezione di una relazione tra due variabili quando, in realtà, tale relazione non esiste. Quando si è già convinti che esista una relazione tra due eventi, si hanno semplicemente maggiori probabilità di notare la loro occorrenza congiunta e, al contrario, meno probabilità di ricordare i momenti in cui viene a mancare. Questo meccanismo è anche alla base delle distorsioni nel comportamento di verifica delle ipotesi: se, ad esempio, in un giorno di pioggia si soffre di emicranie, allora si tende a dare come verificata l'ipotesi per cui la pioggia porta sempre dolori alla testa.

  La correlazione illusoria gioca un ruolo importante anche nella stereotipizzazione: gli stereotipi sono definiti come generalizzazioni riguardanti un gruppo di persone in cui si presume che ogni membro del gruppo abbia bene o male le stesse caratteristiche. Inoltre, quando ad assumere comportamenti negativi è una minoranza, questi rimangono molto impressi nella memoria delle persone poiché più rari, portando erroneamente a credere che i membri appartenenti a quella determinata minoranza abbiano una maggiore probabilità di comportarsi in modo negativo. Questo porta ad una correlazione illusoria tra la rarità del gruppo e la rarità del comportamento.

- L'Effetto senno del poi, infine, è una distorsione che si ha nel giudizio a posteriori di un evento: una volta conosciutone l'esito, si usa credere fosse anche evidentemente il più probabile riguardando i dati disponibili. In sintesi, questo bias avviene quando l'acquisto di una nuova informazione cambia il modo in cui l'intera esperienza viene ricordata, ponendo attenzione soltanto sulle informazioni utili a confermare ciò che si sa a posteriori essere vero, avendo la sensazione di averlo sempre saputo sin dal principio. L'effetto del senno del poi trova le sue radici anche nel concetto di ancoraggio osservato precedentemente: l'esito dell'evento viene utilizzato come un'ancora sul quale regolare i nostri giudizi precedenti.

  Un esempio reale si può osservare, ad esempio, nelle cause contro i medici ospedalieri: la conoscenza a posteriori della diagnosi corretta in

una determinata situazione critica, porta a pensare erroneamente che fosse evidentemente la più probabile, portando spesso i famigliari a credere che la colpa stia nella scarsa preparazione tecnica del medico.

- **Bias di rappresentatività**: Come già osservato precedentemente, il bias di rappresentatività si basa sulla somiglianza o apparenza. Tra i bias di rappresentatività si possono trovare la fallacia della congiunzione, osservata nel Linda problem di Kahneman e Tversky, e la fallacia del giocatore, basata sulla legge dei grandi numeri, osservata invece nell'hospital problem.

- **Effetto carrozzone**: rappresenta una distorsione nel pensiero di gruppo: è un pregiudizio cognitivo che porta a credere a qualcosa semplicemente perché ci credono le altre persone. Può anche portare a credere che qualcosa sia impossibile da realizzare soltanto perché altri hanno già provato e fallito, o che esista un solo metodo per risolvere un problema.
  L'effetto carrozzone è il motivo per cui le recensioni e le valutazioni sottoforma di stelle sono diventate una componente integrante del marketing online: è più probabile che le persone acquistino un prodotto o un servizio se vedono che molte altre persone lo hanno fatto risultandone soddisfatte.

- **Impact bias:** è la tendenza a sopravvalutare l'impatto emotivo di un evento, facendo previsioni sui propri stati emotivi futuri. Questo bias ha un impatto significativo nel processo decisionale su larga e piccola scala. Si potrebbero considerare i sentimenti futuri quando, ad esempio, si decide se sposarsi, avere un figlio o cambiare carriera: le aspettative sull'influenza che queste azioni avranno sulla propria felicità giocheranno quasi certamente un ruolo significativo nella decisione. Gli individui tendono a sopravvalutare la durata e la forza delle emozioni future, aspettandosi che gli eventi attuali abbiano un impatto maggiore sugli stati emotivi futuri di quanto non facciano realmente. Questi errori di previsione sono per lo più prodotti poiché si concentra troppo la propria attenzione su un singolo evento, senza valutare tutti gli altri elementi che influenzeranno. L'impact bias si verifica anche quando nel predire lo stato emotivo futuro non si prende in considerazione la propria flessibilità

psicologica e i metodi di coping[16] utilizzati solitamente per mitigare l'influenza di esiti futuri sfavorevoli. Gli individui spesso dimenticano l'importanza di questo sistema immunitario psicologico, così come ne sottovalutano il potenziale di autoprotezione.

I bias cognitivi favoriscono dunque un alfabetismo funzionale, sostituendo un approccio più razionale e completo con delle risposte intuitive ed immediate. Non esiste un modo per evitare di cadere in questi meccanismi dal momento che nessuno risulta esserne immune. L'unica soluzione è quella di conoscerli e studiarli il più possibile, in modo da riconoscere i contesti di attivazione e mettere in discussione il proprio giudizio.
Negli ultimi anni, in un mondo sempre più digitalizzato, l'individuazione di bias cognitivi all'interno dei dati risulta essere di fondamentale importanza nell'automatizzazione di processi decisionali, diventando così oggetto di studio anche nel campo dell'intelligenza artificiale.

2. Il ruolo dell'Intelligenza Artificiale

2.1 Intelligenza Artificiale: introduzione e storia

Il termine *Intelligenza Artificiale* fu coniato per la prima volta dal matematico John McCarthy nel 1956 e sta ad indicare un'intelligenza creata artificialmente dall'uomo. L'obiettivo dei ricercatori nel campo dell'intelligenza artificiale è quello di imitare l'intelligenza umana attraverso l'utilizzo di algoritmi. La disciplina dell'intelligenza artificiale trova le sue origini negli anni Cinquanta, periodo di grande fermento per le discipline scientifiche. In particolare, quando si parla di origini in fatto di intelligenza artificiale, non si può non pensare alla ricerca[17] svolta da Alan Turing, considerato uno dei padri dell'informatica moderna che, nel 1936, attraverso la

[16] Strategie adattative utilizzate per ridurre le emozioni negative.

[17] A. Turing, *Computing Machinery and Intelligence*, 1950.

macchina di Turing, aveva posto le basi per i concetti di calcolabilità e computabilità. Attraverso il *test di Turing*, infatti, una macchina poteva essere considerata intelligente nel caso in cui il suo comportamento, osservato da un essere umano, potesse essere considerato indistinguibile da quello di una persona. Il test di Turing consisteva in una conversazione, attraverso messaggi scritti, tra la macchina ed un esaminatore per la durata di cinque minuti. Al termine, l'esaminatore doveva indovinare se la conversazione era avvenuta con una macchina o con una persona reale e, se la macchina ingannava l'esaminatore almeno il 30% delle volte, allora il test veniva considerato superato, dal momento che l'umano aveva attribuito alla macchina l'intelligenza di un interlocutore. Il lavoro svolto da Alan Turing portò molta attenzione sulla disciplina da parte della comunità scientifica.

Subito dopo la nascita dell'intelligenza artificiale c'è stato un periodo di entusiasmo precoce dovuto alle grandi aspettative che si erano create tra i ricercatori: l'intelligenza artificiale veniva vista come qualcosa di straordinario date le limitate capacità dei computer dell'epoca, portando alla formazione di grandi aspettative che non si sarebbero poi realizzate per diversi anni. Già nel 1957, Herbert Simon, economista, psicologo e informatico statunitense, suggeriva che nel giro di dieci anni la comunità scientifica sarebbe stata in grado di sviluppare un'intelligenza artificiale in grado di competere con i più grandi campioni di scacchi. Purtroppo, a causa di un'incapacità computazionale adeguata, le aspettative non furono mantenute portando alla frammentazione dell'Intelligenza Artificiale in distinte aree basate su teorie diverse[18].

In questo contesto apparvero due principali paradigmi: l'intelligenza artificiale forte e l'intelligenza artificiale debole:

L'intelligenza artificiale forte è definita come "l'idea che opportune forme di intelligenza artificiale possano veramente ragionare e risolvere problemi(..)che è possibile per le macchine diventare sapienti o coscienti di sé, senza necessariamente mostrare processi di pensiero simili a quelli umani."[19]. Questa corrente di pensiero non

---

[18] *Storia dell'Intelligenza Artificiale, da Touring ai giorni nostri,* maggio 2019, Osservatori.net digital innovation, https://blog.osservatori.net/it_it/storia-intelligenza-artificiale.

[19] *Intelligenza artificiale forte*,10 marzo 2021 ore 23:51, Wikipedia, https://it.wikipedia.org/wiki/Intelligenza_artificiale_forte.

ritiene quindi il computer soltanto uno strumento ma, se programmato opportunatamente, una vera e propria mente.
Al contrario il paradigma dell'intelligenza artificiale debole sostiene che "è possibile sviluppare macchine in grado di risolvere problemi specifici senza avere coscienza delle attività svolte. In altre parole, l'obiettivo(..)non è realizzare macchine dotate di intelligenza umana, ma di avere sistemi in grado di svolgere una o più azioni umane complesse."[20]

A partire dal 1980, l'intelligenza artificiale inizia a diventare un'industria miliardaria che supporta le aziende nella gestione del denaro attraverso l'utilizzo di sistemi complessi e, già dalla metà del 1990, l'intelligenza artificiale diventa un campo multidisciplinare. Il 1997 è un anno di fondamentale importanza per i ricercatori: il computer Deep Blue, costruito dall'azienda IBM per il gioco degli scacchi, batte il campione mondiale nella disciplina Garry Kasparov. Con lo scopo di indagare il significato psicologico e filosofico da attribuire alle capacità della macchina nel simulare processi cognitivi, nel 1980 il filosofo americano John Searle pubblica un esperimento che prende il nome di *Chinese Room*[21]. Nell'esperimento l'autore immagina una situazione simile a quella proposta da Turing: una persona si trova sola all'interno di una stanza e riceve domande da una persona madrelingua cinese presente all'esterno. La persona situata all'interno della stanza possiede diversi simboli cinesi con cui rispondere ma non conosce la lingua cinese, tuttavia, all'interno della stanza, sono presenti delle istruzioni nella sua lingua madre riguardanti i diversi simboli e il modo in cui metterli assieme al fine di rispondere in un determinato modo. Ne risulterà che sarà possibile per la persona all'interno della stanza comunicare con la persona all'esterno in lingua cinese, non perché quella all'interno capisca realmente il cinese o il significato del singolo ideogramma, ma perché comprende le regole, scritte nella propria lingua, che indicano come rispondere utilizzando i simboli presenti. La persona all'interno può quindi dare l'impressione di conoscere il cinese alla persona all'esterno della stanza, ma non ne capisce realmente il significato.

---

[20] *Storia dell'Intelligenza Artificiale, da Touring ai giorni nostri,* maggio 2019, Osservatori.net digital innovation, https://blog.osservatori.net/it_it/storia-intelligenza-artificiale.

[21] In italiano tradotto come "stanza cinese".

Figura 1: esperimento della Stanza Cinese.

Fonte: macrovu.com

Nell'esperimento proposto da Searle la persona presente nella stanza è paragonabile alla macchina nel test di Turing, mentre la persona madrelingua cinese ricopre il ruolo dell'interlocutore. L'autore vuole dimostrare la fallacia del test di Turing e, più in generale, del paradigma dell'intelligenza artificiale forte: nella sua proposta, infatti, il computer utilizza delle regole sintattiche per manipolare simboli o stringe, senza però coglierne il reale significato. Al contrario, la mente umana associa e comprende il reale significato dei simboli. Per questo, lo scopo finale dell'esperimento è quello di dimostrare che l'intelligenza umana e quella della macchina non possono essere considerate equiparabili, processando le informazioni in maniera differente: il ragionamento umano deriva infatti da processi di origine biologica, mentre il computer può soltanto simulare questi processi. Tuttavia, è possibile smuovere diverse critiche all'ideologia portante dell'esperimento, come, ad esempio, il fatto che il significato di comprensione è relativo e che tecniche più avanzate, se inserite all'interno di un corpo robotico, potrebbero replicare il funzionamento dei neuroni all'interno della mente.

Alla fine del 1999 si assiste ad una crescita esponenziale con l'avvento dei *Big Data:* con il termine si fa riferimento ad una grande quantità di dati collezionabili. All'inizio il termine veniva utilizzato con un'accezione quasi negativa, non avendo gli

strumenti adatti per analizzare una simile quantità di dati. Le caratteristiche dei Big Data sono riassumibili in cinque caratteristiche chiave, chiamate anche le cinque *v*:

-Volume: il termine volume fa riferimento alla quantità di dati in proprio possedimento, se è abbastanza grande allora si può parlare di Big Data. Ovviamente la definizione di grandezza varia a seconda del periodo storico in cui ci si trova e delle tecnologie a disposizione. Al momento, società come Amazon, arrivano ad ottenere ogni secondo milioni di dati in tempo reale.

-Velocità: con velocità si fa riferimento alla velocità con cui si accumulano i dati, in modo costante e continuo, da fonti come, ad esempio, social media, telefoni, assistenti vocali ed altro. Questi dati vengono analizzati velocemente, quasi in tempo reale.

-Varietà: il termine fa riferimento alle differenti tipologie di dati. I dati collezionabili possono essere registrati in diversi formati vista la numerosità di fonti esistenti e sono differenziabili principalmente in tre categorie: strutturati, semi strutturati e non strutturati. I dati non strutturati non sono adatti alla tradizionale struttura dei database relazionali, che prevede la loro organizzazione in tabelle, righe e colonne. I dati semi strutturati sono dati che non sono stati organizzati in uno specifico deposito, ma contengono informazioni associate, come, ad esempio, i metadati[22]. Infine, i dati strutturati sono dati organizzati in una determinata struttura di deposito, che ne rende più semplice l'analisi.

-Veridicità: con il termine si va ad indicare il livello di affidabilità dei dati, ovvero la loro qualità e precisione. Avere dei dati realmente affidabili è di fondamentale importanza per ricavarne delle deduzioni sensate, soprattutto nei campi più delicati come quello medico.

-Valore: si fa riferimento al valore che i Big Data possono fornire alle aziende se utilizzati in modo intelligente. Il valore dei Big Data aumenta a seconda delle informazioni e strategie che possono essere ottenute analizzandoli.

[22] Il termine viene utilizzato solitamente in riferimento a dati che ne descrivono altri.

Oggi l'intelligenza artificiale è utilizzata in molteplici campi ed è di fondamentale importanza per lo sviluppo di nuove tecnologie, ma resta ancora incerta la sua reale potenzialità e se potrà mai eguagliare l'intelligenza umana come previsto nella teoria dell'intelligenza artificiale forte. Nonostante ciò, alcuni esperti sostengano possa raggiungere un livello di intelligenza equiparabile a quella umana entro il 2050. Le possibilità che il suo sviluppo porti a conseguenze disastrose per l'umanità sono piuttosto basse, tuttavia, i paesi in grado di sviluppare i migliori algoritmi di intelligenza artificiale hanno un controllo maggiore sull'umanità.

Al giorno d' oggi l'intelligenza artificiale può essere definita come un insieme di teorie, tecniche e discipline che hanno l'obiettivo comune di riprodurre, attraverso una macchina, le abilità cognitive degli esseri umani, in modo da poter delegare alla macchina compiti complessi precedentemente svolti da esseri umani, avendo a disposizione una maggiore velocità di calcolo. Gli algoritmi di intelligenza artificiale, infatti, non sono soltanto in grado di riprodurre ragionamenti umani, ma possono anche essere molto più abili nel farlo poiché in grado di processore migliaia di informazioni in poco tempo.

2.2 Gli algoritmi di apprendimento automatico

Gli algoritmi possono essere considerati il cuore dell'intelligenza artificiale: non solo sono in grado di riprodurre ragionamenti umani, ma possono anche essere molto più abili nel farlo poiché in grado di processore migliaia di informazioni in poco tempo. In particolare, il concetto di algoritmo è alla base del machine learning (o apprendimento automatico), campo di ricerca a cui spesso si fa riferimento con il termine intelligenza artificiale.

Un algoritmo può essere definito come una sequenza di istruzioni finite aventi come obiettivo quello di risolvere un problema specifico. L'algoritmo frammenta il problema in diversi passaggi a seconda della sua complessità, ovviamente più passaggi sono presenti in un algoritmo, più questo può essere definito complesso. Solitamente gli algoritmi vengono scritti attraverso appositi linguaggi di programmazione come, ad esempio, Python. Il ruolo dell'algoritmo nel campo dell'intelligenza artificiale può essere spiegato velocemente in questo modo: quando si colleziona una grande quantità di dati, generata da diversi utenti e apparecchi, questa quantità viene data in input ad

un algoritmo scritto in modo da frammentare il problema in problemi minori e, infine, generare un output che consiste nella soluzione desiderata.

Il campo specifico del machine learning studia e implementa, attraverso nozioni di statistica, algoritmi (o modelli) di intelligenza artificiale da alimentare con dati strutturati e suddivide gli algoritmi in due categorie principali: algoritmi di apprendimento supervisionato e algoritmi di apprendimento non supervisionato.

Nell'apprendimento supervisionato, i modelli vengono addestrati utilizzando dati etichettati, ciò significa che i dati in entrata sono suddivisi in diverse categorie ben definite, tra cui la classe. Con il termine *classe* si fa riferimento alla categoria target, ovvero ciò che si vuole prevedere attraverso l'utilizzo dell'algoritmo, e di cui le altre categorie ne rappresentano le caratteristiche (o features). In questa forma di apprendimento, quindi, l'algoritmo utilizza questi dati con un obiettivo ben preciso. Per fare in modo che l'algoritmo capisca come prevedere la classe precedentemente scelta, si dividono i dati in due parti: la prima parte, detta anche *training set*, conterrà tutte le categorie, inclusa la classe, e verrà utilizzata dall'algoritmo per allenarsi. La seconda parte di dati, detta anche *test set*, conterrà invece tutte le categorie eccetto la classe da predire. Durante l'allenamento, in cui si utilizza quindi la prima parte di dati di cui si conosce la classe di arrivo, l'algoritmo dona a ciascuna feature un peso diverso in base all'influenza che essa ha sulla classe. In questo modo, quando verrà data in input la seconda parte di dati non contenente la classe, l'algoritmo utilizzerà le features e i loro relativi pesi generati precedentemente per predirne la classe di arrivo.

L'apprendimento supervisionato è utilizzato in molteplici campi. Un esempio ricorrente e semplice da comprendere è quello dei prestiti finanziari: l'algoritmo viene allenato su dati etichettati come, ad esempio, l'età della persona richiedente il prestito, il suo contratto di lavoro e il suo guadagno annuo al fine di predire la classe di arrivo, ovvero se erogare o meno il prestito. Questi dati servono a comprendere quanto sia probabile che l'individuo riesca o meno a ripagare il debito.

Nell'apprendimento non supervisionato, invece, non ci sono risposte corrette o classi di arrivo prestabilite. L'algoritmo apprende e forma dei raggruppamenti direttamente dai dati in input, senza bisogno di allenarsi precedentemente. Per dare una forma, ovvero l'output, a questi dati, solitamente si definisce una metrica con cui l'algoritmo calcolerà la distanza tra i punti**.** I punti non sono altro che i dati in input distribuiti e osservati all'interno di uno spazio. Questa distanza può essere osservata in diversi modi: si può prendere in considerazione la distanza tra un punto e il suo punto più

vicino come la distanza tra un punto e un insieme di punti. Il motivo per cui si utilizza la distanza per dare forma a questi dati e raggrupparli è perché, in questo caso, la distanza definisce la somiglianza tra i diversi dati presi in input: più due punti (o dati) sono vicini nello spazio, più sono simili tra loro.
Gli algoritmi di apprendimento non supervisionato sono spesso utilizzati, ad esempio, nei sistemi di raccomandazione di Netflix o Spotify: l'algoritmo prende in considerazione la distanza, ovvero la somiglianza, tra due punti che, in questo caso, possono rappresentare un film, una serie o una canzone.

Come già accennato precedentemente, negli ultimi anni la quantità di dati reperibili è aumentata vertiginosamente dando vita a quelli che oggi vengono definiti Big Data. Al fine di processare enormi quantità di dati, negli ultimi anni, è diventato di fondamentale importanza il campo del deep learning, o apprendimento profondo. Il deep learning è un sottoinsieme del machine learning che comprende algoritmi molto complessi come, ad esempio, le reti neurali con tre o più livelli. L'obiettivo è quello di simulare il comportamento del cervello umano, permettendo l'analisi di grandi quantità di dati. Tuttavia, anche se si prendessero in considerazione algoritmi molto complessi come quelli di deep learning, le loro performance non sarebbero utili o attendibili se provate soltanto su una piccola quantità di dati. In breve, nonostante avere un algoritmo ben strutturato sia importante, la quantità di dati su cui questo viene allenato lo è ancora di più. Questo accade perché, se l'algoritmo viene allenato soltanto su una piccola quantità di dati, non è possibile sapere quale sia in generale il suo reale livello di performance[23]: l'algoritmo potrebbe infatti dare buoni risultati nella predizione di quella piccola quantità di dati, risultando però mediocre nella predizione di dati differenti, poiché allenatosi soltanto su poche informazioni. Inoltre, basta pensare alla legge dei grandi numeri: più spesso viene ripetuta un'attività o, in questo caso, un allenamento, maggiori sono le possibilità che si raggiunga un livello costante nella performance. A sostegno di questa tesi c'è anche una famosa citazione di Peter Norvig, capo del dipartimento scientifico di Google, che descrive in questo modo il segreto del successo dell'azienda:

[23] Prestazione.

"*We don't have better algorithms than anyone else; we just have more data"*[24]

L'apprendimento automatico apre quindi le porte ad un'infinità di campi ed impieghi. Le macchine oggi possono imparare la punteggiatura, la grammatica e creare testi completamente nuovi e corretti. Inoltre, sono in grado di identificare e creare immagini, proprio come un essere umano. Ci sono modelli di deep learning specializzati in segnali stradali, altri addestrati a riconoscere i pedoni lavorando contemporaneamente alla guida autonoma di un'automobile. Ormai l'intelligenza artificiale è utilizzata in quasi ogni settore, con particolare rilevanza in alcuni di essi.

2.3 Principali settori di impiego

Nel corso degli ultimi anni è diventata di maggiore necessità la digitalizzazione di diversi settori. L'evoluzione delle tecnologie ha permesso la trasformazione di numerosi servizi, in modo da facilitare parte della burocrazia e migliorare la vita delle persone, permettendo così di svolgere diverse attività, come ad esempio la ricerca di un alloggio o l'attivazione di un servizio postale, comodamente da casa. Tuttavia, come ogni forma di evoluzione, anche questa ha lasciato indietro un gran numero di imprese o persone che non hanno saputo adattarsi a causa, ad esempio, della loro età o del tipo di industria, risultando per loro in una difficoltà più che in un'agevolazione. Di pari passo con l'evoluzione tecnologica c'è anche l'avvento dei social media, ormai fonte non solo di intrattenimento, ma anche di lavoro per numerose persone.
Sarebbe impossibile elencare tutti i settori in cui l'intelligenza artificiale svolge, ad oggi, un ruolo fondamentale; tuttavia, di seguito, verranno introdotti alcuni tra i principali campi di impiego, in modo da mostrare una panoramica delle sue possibili applicazioni:

-Pubblica amministrazione: come già accennato precedentemente, l'intelligenza artificiale è stata di fondamentale importanza nella pubblica amministrazione. Un

[24] [Mia traduzione] "Non abbiamo algoritmi migliori degli altri, abbiamo soltanto più dati"

esempio evidente della digitalizzazione dei servizi è la diffusione dell'identità digitale, necessaria per accedere ad un elevato numero di servizi senza dover mostrare o inserire documenti. Attraverso l'utilizzo di questi servizi è diventato possibile richiedere ed ottenere documenti senza la necessità di recarsi in sede, riducendo così i costi del personale, l'accumulo di persone agli sportelli e l'archivio di un numero elevato di documenti cartacei.

-Intrattenimento: al giorno d'oggi, sono presenti miliardi di profili su piattaforme come Instagram, Twitter e Facebook, che fanno affidamento ad algoritmi di intelligenza artificiale per organizzare e monitorare enormi quantità di dati.
L'utilizzo dell'intelligenza artificiale all'interno dei social media è presente in diverse forme. Nella maggior parte dei social è possibile trovare il riconoscimento di immagini: alcuni social, infatti, spesso riconoscono quando è presente all'interno di una foto il volto di una persona. Altri esempi del suo impiego all'interno dei social media possono essere il suggerimento di ricerche in base alle proprie ricerche più recenti, o i contenuti suggeriti all'interno del proprio feed in base alle prefernze mostrate dall'utente. Per quanto riguarda la pubblicazione e la moderazione di contenuti non appropriati, anche qui solitamente si trova una combinazione di moderatori umani e algoritmi di intelligenza artificiale in grado di riconoscere attraverso immagini o parole se il contenuto sia o meno da rimuovere.
Anche gli assistenti vocali possono essere considerati un ulteriore forma di intrattenimento: essi, infatti, non sono soltanto utili per effettuare ricerche, cambiare il colore delle luci nella propria casa o scrivere una lista, ma possono anche essere utilizzati per ascoltare la musica, giocare a dei quiz e dialogare. Tra i più famosi ci sono l'assistente vocale di Apple, che prende il nome di Siri, e quello costruito da Amazon che prende il nome di Alexa. Questi assistenti utilizzano al loro interno algoritmi avanzati di elaborazione del linguaggio naturale in modo da comprendere ciò che viene detto dall'utente e formulare una risposta parlata.
Applicazioni come Vinted e Shein[25], invece, utilizzano l'intelligenza artificiale per la costruzione di chatbot, ovvero agenti virtuali con cui dialogare in caso di problemi, prima di ricorrere all'intervento umano. In questo modo è possibile

[25] Applicazioni per la vendita di beni materiali come vestiti, oggetti per la casa e altro.

risparmiare tempo e denaro non dovendo impiegare da subito risorse umane per assistere gli utenti in caso di piccoli dubbi.
Inoltre, nell'ultimo anno, ha preso sempre più piede l'utilizzo di strumenti di intelligenza artificiale per la creazione di illustrazioni o grafiche: in questi casi, l'intelligenza artificiale sostituisce la capacità di disegno dell'uomo che deve, invece, essere abile nello scegliere i giusti parametri da dare in input alla macchina per creare il prodotto desiderato.

-Trasporti: Sono molteplici gli utilizzi dell'intelligenza artificiale nel settore dei trasporti. L'utilizzo più attuale ed innovativo è quello delle macchine con pilota automatico: queste macchine sono dotate di sensori che raccolgono dati ininterrottamente. Solitamente queste macchine sono in grado di collezionare dati riguardanti la velocità della macchina, la presenza di ulteriori macchine attorno ad essa, la presenza di pedoni e le condizioni e la forma della strada in cui si trova. Successivamente, attraverso algoritmi di intelligenza artificiale che analizzano costantemente questi dati, la macchina riesce a prendere decisioni in base al contesto in cui si trova.
Ulteriori impieghi di algoritmi di apprendimento automatico nell'ambito dei trasporti si possono trovare nelle applicazioni di navigazione, come ad esempio Google Maps o Waze, che consentono all'utente di conoscere, in tempo reale, la strada migliore per spostarsi da un posto ad un altro, in macchina, in bicicletta o a piedi. Google Maps utilizza, infatti, algoritmi di intelligenza artificiale per scannerizzare informazioni relative alle condizioni della strada e del traffico.

-Finanza: All'interno del settore finanziario gli algoritmi di intelligenza artificiale sono diventati di fondamentale importanza per la comprensione dell'andamento economico in diversi paesi. Questo settore genera un numero elevato di dati, di cui l'analisi risulta essere fondamentale per le banche. Gli algoritmi di apprendimento automatico sono in grado di semplificare sfide finanziarie come la gestione dei rischi o dei prestiti: le aziende, infatti, al fine di ottenere una crescita rapida e solida, hanno bisogno di previsioni sempre più accurate. Per questo, identificare modelli di comportamenti passati che hanno portato a incidenti e trattarli conseguentemente come un rischio, risulta essere di fondamentale importanza. Anche per quanto riguarda i prestiti, come già accennato precedentemente, l'intelligenza artificiale

permette di comprendere quali persone possono o meno ottenere un prestito e quanto elevato, in modo da avere una stima della possibilità che queste persone hanno di ripagarlo nel corso degli anni o di chiederne di ulteriori.

Inoltre, l'utilizzo dell'intelligenza artificiale permette agli investitori privati di prendere decisioni lungimiranti sui loro investimenti, traendo vantaggi dai mercati in evoluzione: è possibile, infatti, generare consulenze e consigli personalizzati per la gestione del proprio patrimonio. Infine, un'ulteriore applicazione in campo finanziario, importante dal punto di vista della legalità, è la prevenzione di frodi: l'intelligenza artificiale, infatti, è in grado di analizzare i dati in modo da estrapolarne le anomalie che potrebbero, invece, passare inosservate all'essere umano. L' individuazione di queste irregolarità non è fondamentale soltanto per evitare attività di riciclaggio di denaro o attacchi informatici, ma anche per creare un rapporto di fiducia con i clienti che, ovviamente, desiderano che il loro denaro sia al sicuro sia all'interno del conto che durante i pagamenti online.

-Sanità: L'intelligenza artificiale offre un supporto fondamentale alla sanità, già da ancor prima dell'avvento del covid-19. La ricerca in questo campo si occupa di studiare metodi utili per analizzare grandi quantità di dati, in modo da migliorare diagnosi e prevenzione. Oggi, gli algoritmi stanno già superando il lavoro dei radiologi nell'individuazione dei tumori maligni, guidando i ricercatori sul modo in cui proseguire per quanto riguarda gli studi clinici. Le principali categorie di applicazione riguardano la diagnosi e le raccomandazioni terapeutiche, il coinvolgimento e l'aderenza del paziente e le attività amministrative. Sebbene ci siano molti casi in cui l'intelligenza artificiale può svolgere compiti sanitari meglio degli umani, i fattori di implementazione impediscono l'automazione su larga scala: oltre ai costi elevati, vi sono da considerare anche i possibili problemi etici che verrebbero a crearsi. Nel settore sanitario, l'applicazione più comune dell'apprendimento automatico tradizionale è la medicina di precisione, che riesce a stabilire quali protocolli di trattamento possono avere successo o meno su un paziente basandosi sulle sue caratteristiche e sul contesto del trattamento. La grande maggioranza delle applicazioni di apprendimento automatico nella medicina di precisione richiede un apprendimento supervisionato, ovvero la formazione di un set di dati di cui è nota la variabile di esito. Un'applicazione comune del deep learning è il riconoscimento di lesioni potenzialmente cancerose nelle immagini

radiologiche. Il deep learning viene applicato sempre più spesso per il rilevamento di caratteristiche clinicamente rilevanti nei dati di imaging[26] , andando oltre ciò che può essere percepito dall'occhio umano.

L'utilizzo dell'apprendimento automatico viene utilizzato, in questo campo, anche per prevedere le possibili popolazioni a rischio di particolari malattie o incidenti.

Alcune organizzazioni sanitarie hanno anche sperimentato chatbot per l'interazione con il paziente: queste applicazioni possono essere utili per scopi semplici come la fissazione di appuntamenti. Tuttavia, in un sondaggio condotto su 500 pazienti statunitensi riguardante i primi cinque chatbot utilizzati nel settore sanitario, i pazienti hanno espresso preoccupazione per la rivelazione di informazioni riservate[27].

Infine, i robot chirurgici, inizialmente approvati negli Stati Uniti nel 2000, migliorano la capacità di vedere, creare incisioni precise e minimamente invasive, suturare ferite e così via. Tuttavia, le decisioni importanti sono ancora prese dai chirurghi umani.

-Giudiziario: in questo campo, l'intelligenza artificiale, è utilizzata per la valutazione del rischio e consiste in algoritmi che utilizzano i dati degli imputati per analizzare il rischio di recidività, ovvero la probabilità che ripetano il crimine in futuro. Inoltre, l'utilizzo dell'intelligenza artificiale per il riconoscimento facciale in ambito giudiziario ha ricevuto molte critiche: affermare che le caratteristiche fisiche rappresentino in qualche modo delle caratteristiche personali interiori può essere molto pericoloso.

-Logistica e Distribuzione: nel settore della logistica e della distribuzione, soprattutto nelle grandi aziende, l'intelligenza artificiale ha permesso una maggiore velocità ed efficienza, riducendo così anche i costi. Attraverso gli algoritmi di apprendimento automatico è possibile, infatti, coordinare miliardi di singoli prodotti e merci nel processo di spedizione, che spesso si estende globalmente. Gli ulteriori impieghi in questo settore sono molteplici: i sistemi di intelligenza

[26] Immagini di carattere scientifico, specificatamente a scopo diagnostico.

[27] T. Davenport, *The potential for artificial intelligence in healthcare*, giugno 2019, PMC, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/

artificiale permettono, infatti, una migliore coordinazione di domanda e offerta, la movimentazione delle merci attraverso degli apparecchi robotici, la pianificazione della manutenzione dei veicoli e molto altro ancora. Tutto ciò ha permesso di raggiungere un'ottimizzazione nella produzione, logistica, stoccaggio e consegna, prima inimmaginabili.

-Agroalimentare: nel settore agroalimentare, le nuove tecnologie, hanno contribuito alla sostenibilità e qualità della produzione, rendendo inoltre possibile la diminuzione dei costi. Grazie all'intelligenza artificiale è stato possibile automatizzare interi processi, dalla produzione agricola alla distribuzione finale.
Per quanto riguarda la sostenibilità, di fondamentale importanza negli ultimi anni, l'intelligenza artificiale permette di ridurre l'impatto ecologico della produzione di alimenti, analizzando i dati provenienti da tutta la filiera. Anche la gestione dei campi viene digitalizzata, grazie ad algoritmi che analizzano dati di sensori di rilevamento della temperatura, dell'umidità, dei livelli di irraggiamento e altro, rendendo così possibile l'elaborazione di strategie relative alla gestione del terreno, all'organizzazione dell'irrigazione in modo da evitare sprechi e alla distribuzione della concimazione. Il monitoraggio costante di questi dati, inoltre, permette di creare modelli statistici per l'analisi preventiva e predittiva dell'andamento delle condizioni del suolo, fondamentale per ottenere un'alta qualità nel prodotto finale.
L'intelligenza artificiale, in questo campo, risulta essere utile anche per quanto riguarda le strategie aziendali, fornendo a chi di dovere strumenti di analisi rapidi ed accurati. I dati collezionati sono utilissimi al fine di comprendere e tracciare le abitudini, le esigenze e gli interessi dei consumatori., permettendo così di prevedere i bisogni dei clienti e fornire soluzioni adeguate.
Infine, un ulteriore impiego degli algoritmi di apprendimento automatico si trova nel processo di confezionamento del prodotto finito: questa tipologia di compiti è estremamente ripetitiva e monotona e, per questo, si presta perfettamente all'automatizzazione.

-Beni culturali: negli ultimi anni ha acquisito una grande importanza la digitalizzazione del patrimonio culturale del proprio paese, poiché permette di esplorare nel dettaglio siti di interesse archeologico, riproducendoli sottoforma di modelli virtuali. In questo settore, lo sviluppo e l'utilizzo della robotica ha permesso

di esplorare luoghi difficili da raggiungere e pericolosi per l'essere umano come catacombe, cave, bunker o luoghi sommersi dalla natura o dalle macerie. Attraverso l'utilizzo di dispositivi leggeri è stato possibile sostituire l'attività degli archeologi all'interno di questi siti.
Inoltre, la robotica viene utilizzata anche all'interno di musei per migliorare l'esperienza dei visitatori: molti musei, infatti, mettono a disposizione dispositivi per la realtà aumentata, in modo da rendere più immersiva l'esperienza. Alcuni esempi di queste applicazioni si possono trovare, ad esempio, nella visita guidata a casa Batllò situata nella città di Barcellona, dove è possibile effettuare un tour all'interno del palazzo accompagnati da una video guida in grado di combinare la struttura dell'edificio con elementi virtuali ed animazioni, o al Circo Massimo a Roma, dove, attraverso l'utilizzo di visori per la realtà aumentata, i visitatori possono immergersi all'interno di scenari molto antichi.

-Aerospaziale: nel settore aerospaziale l'intelligenza artificiale ricopre molteplici ruoli. Tra gli eventi più recenti ed importanti per il progresso scientifico vi è la realizzazione della prima foto di un buco nero, destinata a segnare un'evoluzione epocale nella storia dell'astronomia. L'immagine è stata creata attraverso algoritmi di computer vision[28], allenati attraverso diverse immagini spaziali: in questo modo gli algoritmi sono stati in grado di imparare come sono fatti, tipicamente, gli elementi e i corpi presenti nell'universo. Questi algoritmi, combinati a dati raccolti da otto diversi telescopi in tutto il mondo, hanno permesso la realizzazione dell'immagine nel 2019.

L'intelligenza artificiale, accompagnata da una collezione di dati ininterrotta, ricopre dunque un ruolo ormai imprescindibile in molteplici campi all'interno della società. Tuttavia, il progresso tecnologico non porta con sé soltanto i vantaggi e le comodità relative all'automatizzazione dei processi, ma anche preoccupazioni e problemi relativi alla privacy degli utenti e agli errori che la macchina è in grado di commettere se non regolata in modo opportuno.

[28] Tecniche che rendono possibile, attraverso l'utilizzo del computer, la riproduzione di funzioni e processi dell'apparato visivo umano.

2.4 Rischi e regolamenti

Considerando il potenziale e l'attuale utilizzo dell'intelligenza artificiale, appare evidente la necessità di un'adeguata regolamentazione in grado di contenere i rischi che si possono formare. Gli algoritmi di apprendimento automatico non sono sempre perfetti e possono, talvolta, sbagliare, danneggiando gravemente sia il singolo individuo che la società: la moderazione all'interno dei social media potrebbe ostacolare la libertà di parola se utilizzata inappropriatamente, un errore di calcolo in una macchina con l'autopilota potrebbe creare un incidente o si potrebbero creare delle discriminazioni in diversi campi a seconda dei dati che vengono dati in input all'algoritmo per imparare a trarre delle conclusioni. L'intelligenza artificiale è utilizzata anche per prendere decisioni che possono avere un grande impatto sulla popolazione, per questo è importante poter fare riferimento ad un'adeguata regolamentazione in grado di stabilire delle responsabilità di fronte ad un possibile errore della macchina. Già da diversi anni, governi e ricercatori hanno tentato di introdurre delle normative con linee guida per affrontare i rischi e proporre strategie relative all'implementazione di algoritmi di apprendimento automatico all'interno delle varie organizzazioni. Tuttavia, nonostante questi documenti siano utili al fine di attirare l'attenzione sulla tematica, vi è una mancanza di vere e proprie regole istituzionali in grado di regolamentare il sistema, poiché le regole esistenti sono spesso di natura vaga e rendono di conseguenza difficile stabilire una direzione pratica nel loro implemento. Nel corso dell'ultimo anno, ci si sta concentrando nella redazione di un nuovo possibile testo che regoli opportunamente l'intelligenza artificiale, obbligando i fornitori ad informare immediatamente i distributori e qualsiasi altro attore coinvolto di qualsiasi non conformità rilevata sul sistema e delle azioni correttive intraprese per rimediare. Ad oggi, nella maggior parte dei casi, i sistemi di intelligenza artificiale sono regolati da altre normative esistenti, tra cui la protezione dei dati, la protezione dei consumatori e le leggi sulla concorrenza del mercato. Ne risulta che maggiori sono i rischi previsti dall'utilizzo di algoritmi in determinati contesti, maggiori saranno gli obblighi riguardanti la trasparenza relativi al funzionamento dell'algoritmo.

2.4.1 L'importanza della trasparenza

La trasparenza è uno dei presupposti fondamentali per il corretto trattamento dei dati, tutelato dal GDPR[29], ovvero il regolamento generale sulla protezione dei dati. Secondo il GDPR, infatti, ogni utente che utilizza un'applicazione o altro, deve essere a conoscenza di quali dati personali saranno oggetto di trattamento e deve, inoltre, essere indicata la finalità e la durata del trattamento in questione, garantendo così all'utente il controllo sui propri dati.

L'argomento della trasparenza viene introdotto nell'articolo 5 del GDPR che afferma: "*i dati personali sono trattati in modo lecito, corretto e trasparente nei confronti dell'interessato*". Una trasparenza nel trattamento dei dati personale prevede, dunque, che sia presente un'informativa per gli utenti riguardante il trattamento, che sia possibile conoscere inoltre le modalità in cui il trattamento viene o meno consentito in modo che l'utente possa esercitare il proprio diritto. L'informativa sul trattamento dei dati personali è quindi la massima espressione della trasparenza, per questo è necessario che l'utente accetti sempre il trattamento prima di poter utilizzare un'applicazione o accedere ad un sito. Di conseguenza, è importante che l'utente sia in grado di comprendere l'informativa proposta ed avere una totale comprensione di ciò che sta leggendo ed accettando: il concetto di trasparenza, infatti, si riflette anche nella comunicazione all'utente, la quale deve essere scritta con un linguaggio chiaro e semplice, in forma concisa e facilmente accessibile. A tal proposito si esprime l'articolo 12 del GDPR, che afferma quanto segue: "*Il titolare del trattamento adotta misure appropriate per fornire all'interessato tutte le informazioni di cui agli articoli 13 e 14 e le comunicazioni di cui agli articoli da 15 a 22 e all'articolo 34 relative al trattamento in forma concisa, trasparente, intelligibile e facilmente accessibile, con un linguaggio semplice e chiaro, in particolare nel caso di informazioni destinate specificamente ai minori. Le informazioni sono fornite per iscritto o con altri mezzi, anche, se del caso, con mezzi elettronici. Se richiesto dall'interessato, le informazioni possono essere fornite oralmente, purché sia comprovata con altri mezzi l'identità dell'interessato".* La previa informazione riguardante il trattamento dei propri dati consente, inoltre, di evitare un numero eccessivo di richieste da parte degli utenti interessati a conoscere le modalità, la quantità e la tempistica in cui i loro dati vengono

[29] General Data Protection Regulation.

trattati. Nonostante ciò, è comunque possibile richiedere ed ottenere informazioni maggiori al riguardo in modo, solitamente, gratuito.
Il concetto di trasparenza ha assunto forme sempre più complesse negli ultimi anni: in molti algoritmi di intelligenza artificiale come, ad esempio, le reti neurali, non è sempre chiara la modalità in cui l'algoritmo prende le decisioni al suo interno. Questi algoritmi vengono definiti algoritmi a scatola nera, o *black box*, e rendono complicato spiegare il motivo per cui la macchina fornisce un output piuttosto che un altro. Un esempio pratico di questa problematica si può trovare facilmente in ambito medico: se un paziente viene informato che l'analisi di un'immagine relativa al suo corpo ha portato ad una determinata diagnosi, probabilmente vorrà conoscerne i motivi. In questi casi i medici potrebbero non essere in grado di fornire una spiegazione adeguata. Tuttavia, nel corso degli ultimi anni, sono state sviluppate diverse tipologie di algoritmi in grado di "sbirciare" all'interno della scatola nera di questi algoritmi, in modo da comprenderne, almeno in parte, il loro comportamento interno.

2.4.2 Privacy nella raccolta dati

La privacy consiste nella riservatezza delle informazioni personali. Nell'ambito dell'intelligenza artificiale, si utilizzano spesso i termini *data security* e *data privacy* che consistono in due concetti tra loro collegati. Con il termine *data security*, o sicurezza dei dati, si fa riferimento ai processi necessari per la protezione delle informazioni personali da terze parti non autorizzate, ad eccezione, a volte, di possibili attacchi dannosi esterni. Il termine *data privacy*, invece, fa riferimento ai processi necessari al fine di garantire un adeguato utilizzo dei dati personali forniti dagli utenti, informando il modo in cui questi sono collezionati, utilizzati e condivisi. Oggi, la raccolta dei dati è di fondamentale importanza per l'economia, poiché i dati vengono utilizzati, spesso, per influenzare la popolazione attraverso strategie di marketing: il successo di Google, ad esempio, risiede anche nel modo in cui colleziona i dati degli utenti: chiunque utilizza il motore di ricerca, infatti, non può nascondere i propri interessi quando ricerca informazioni in proposito. Un altro metodo, utile al fine di collezionare un grande numero di dati, è quello di raccoglierli attraverso una sorta di sorveglianza continua, come accade attraverso i dispositivi Amazon Alexa e Google Home, assistenti vocali che registrano continuamente i dati audio provenienti dalle abitazioni private per poi depositarli, collezionarli e analizzarli. Le persone sembrano

avere idee contrastanti quando si fa riferimento alla privacy dei loro dati: la maggior parte degli utenti, infatti, si mostra solitamente molto preoccupata sulla questione. Tuttavia, gli stessi utenti, sembrano accettare facilmente l'invasione della propria privacy quando possono trarne dei benefici. Nel corso della storia ci sono stati importanti casi di violazione della privacy: nel 2013 l'americano Edward Joseph Snowden ha rivelato pubblicamente dettagli di programmi segreti di sorveglianza di massa, che vedevano coinvolta, in particolare, la National Security Agency[30].
Uno dei problemi maggiori riguardante la privacy dei dati è la poca attenzione da parte degli utenti: la maggior parte, infatti, accetta i termini e le condizioni di utilizzo di un sito o di un'applicazione senza neanche leggerli e questo comportamento meccanico, se normalizzato, può portare a serie conseguenze.

2.4.3 La forza lavoro

Ad accompagnare l'evoluzione tecnologica vi è una notevole preoccupazione riguardante l'automazione dei posti di lavoro che porterebbe, di conseguenza, al sostanziale spostamento della forza lavoro. Una collaborazione di Deloitte con l'Oxford Martin Institute ha suggerito che il 35% dei posti di lavoro nel Regno Unito potrebbe essere eliminato dall'intelligenza artificiale nei prossimi 10-20 anni.
Tuttavia, finora, la difficoltà riscontrata nell'integrazione dell'intelligenza artificiale nei flussi di lavoro, soprattutto in ambito clinico, è stata in qualche modo responsabile della mancanza di un vero e proprio impatto sul lavoro. In ambito sanitario, i lavori con maggiori probabilità di essere automatizzati sembrano essere quelli che implicano la gestione di informazioni digitali, radiologia e patologia, piuttosto che quelli con il contatto diretto con il paziente. Ma, anche in lavori come radiologo e patologo, è probabile che la penetrazione dell'intelligenza artificiale in questi campi sia lenta: i radiologi, infatti, fanno di più che leggere e interpretare immagini, mentre i modelli di deep learning sono addestrati soltanto per attività specifiche di riconoscimento delle immagini. Un ulteriore ostacolo in ambito sanitario consiste nella mancanza di dati: gli algoritmi di deep learning utilizzati per il riconoscimento delle immagini necessitano di un approccio supervisionato e, quindi, di "dati etichettati", ovvero

[30] Agenzia per la sicurezza nazionale degli Stati Uniti d'America.

milioni di immagini di pazienti che hanno ricevuto una diagnosi definitiva di cancro, frattura di un osso o altra patologia. Tuttavia, non esiste un archivio aggregato di immagini radiologiche, etichettate o meno.
Infine, vi è ancora una certa riluttanza da parte di alcuni medici e pazienti nel suo impiego. Questa sfiducia ha origine dalla difficoltà nel comprendere il modo in cui la macchina prende decisioni e dalla sua mancanza di empatia.[31]

2.4.4 Pregiudizi e discriminazione

L'Intelligenza Artificiale comporta anche il rischio di pregiudizi, e quindi di discriminazione, all'interno di algoritmi e dati. Per questo, è fondamentale che i produttori siano consapevoli dei rischi e riducano al minimo i potenziali pregiudizi in ogni fase del processo di sviluppo degli algoritmi. Diversi esempi hanno dimostrato che gli algoritmi possono mostrare pregiudizi provocando ingiustizie per quanto riguarda le origini etniche, il colore della pelle, il genere, l'età o la disabilità. Le spiegazioni di tali pregiudizi sono diverse e possono essere sfaccettate: possono, ad esempio, derivare dai set di dati stessi, dal modo in cui i data scientist e i sistemi di machine learning scelgono e analizzano i dati e dal contesto in cui questi vengono utilizzati. Ad esempio, se si immagina un software di supporto alle decisioni cliniche basato sull'intelligenza artificiale che aiuta i medici a trovare il miglior trattamento per i pazienti con cancro della pelle, se l'algoritmo al suo interno è stato prevalentemente addestrato su pazienti caucasici, il software fornirà probabilmente raccomandazioni meno accurate o addirittura imprecise per le sottopopolazioni per le quali i dati di addestramento risultano essere poco inclusivi, come gli afroamericani.
Di norma, ci si aspetta che le decisioni prese dagli algoritmi di apprendimento automatico siano sempre oggettive e neutrali, dal momento che si basano sui dati raccolti e non hanno di per sé emozioni, spesso causa di errori nel giudizio umano. Tuttavia, anche gli algoritmi possono fornire output discriminatori, contenenti quindi un bias, ovvero un errore nella rappresentazione oggettiva dei fatti. Questo errore

[31] *AI and Empathy*: *Combining artificial intelligence with human ethics for better engagement*, luglio 2019, Pega, https://www.pega.com/system/files/resources/2019-11/pega-ai-empathy-study.pdf

accade perché i pregiudizi umani vengono trasferiti alla macchina attraverso i dati con cui viene allenato l'algoritmo: questi dati, infatti, potrebbero rappresentare solo parte della popolazione, ereditare pregiudizi e processi cognitivi errati o riflettere semplicemente i pregiudizi esistenti nel mondo, portando ad una rappresentazione errata della realtà. Il bias, inoltre, può essere causato da un'eccessiva semplificazione della realtà da parte dell'algoritmo, dal momento che i modelli di intelligenza artificiale possono avere maggior difficoltà ad interpretare la totale complessità dei dati.

3 La macchina discriminatoria

3.1 Bias all'interno di algoritmi e dati

Come osservato nel capitolo precedente, i bias non sono presenti soltanto nell'uomo, ma anche all'interno della macchina. La differenza tra i bias presenti nell'uomo e quelli presenti nella macchina deriva, innanzitutto, dalla loro origine: i bias cognitivi presenti nell'uomo, infatti, hanno un'origine naturale dal momento che trovano le loro radici in processi cognitivi inconsci e istinti biologici. Tuttavia, il fatto che la loro origine sia naturale non significa che sia giusto introdurli all'interno della macchina. L'introduzione involontaria di questi errori all'interno di algoritmi di apprendimento automatico può derivare da numerose cause: in alcuni casi, ad esempio, i dati possono non essere rappresentativi di alcuni gruppi esistenti, questo fenomeno può accadere per diversi motivi come la mancanza di accesso a determinate tecnologie, utili a collezionare dati, da parte di paesi generalmente più poveri. Inoltre, nel processo di collezione dei dati, sono gli individui a scegliere il modo in cui collezionarli e quali: un'inadeguata metodologia di raccoglimento può portare ad una mancanza di dati rilevanti per rappresentare determinati gruppi. La mancanza di dati rappresentativi non è l'unico modo in cui i bias entrano a far parte degli algoritmi: i dati collezionati, infatti, seppur presenti in modo adeguato, possono riflettere i pregiudizi esistenti nella società rendendo così la macchina non oggettiva.

In breve, i bias all'interno dell'intelligenza artificiale avvengono quando i risultati prodotti non generalizzano in maniera sufficiente la realtà, e possono essere introdotti

dalla modalità di raccoglimento dei dati, dal modo in cui sono strutturati gli algoritmi e dal modo in cui vengono successivamente interpretati i dati.
Dal momento che tutti i dati contengono bias al giorno d'oggi, è necessario stimare un margine di errore attorno ad ogni dato nell'interpretazione dei risultati, questo errore si riflette nella misura della *confidence*: più l'algoritmo è sicuro dell'appartenenza di un dato ad una determinata classe, maggiore sarà la confidence.
Di conseguenza, per quanto sia entusiasmante l'idea di un'intelligenza in grado di automatizzare attività umane, la storia mostra che le macchine sono efficienti quanto lo è l'essere umano e, quindi, se un dataset conterrà del bias, allora l'intelligenza artificiale lo apprenderà.

Il fatto che i bias cognitivi umani si riflettano all'interno degli algoritmi di apprendimento automatico, risultando in possibili discriminazioni, non è così sorprendente una volta compresi i meccanismi per cui questo fenomeno avviene. Meno ovvio è, invece, il fatto che l'output finale non rifletta soltanto il bias iniziale, ma lo amplifichi. I motivi per cui ciò accade sono da ricercare nel modo in cui l'algoritmo classifica i dati nelle relative classi: nel processo di classificazione, solitamente, gli algoritmi tendono a massimizzare il livello di accuratezza, ovvero quanto l'output finale è classificato correttamente e, per farlo, spesso il modello prenderà in maggiormente in considerazione determinate informazioni piuttosto che altre, amplificando il bias. Ne risulta, dunque, che modelli molto complessi e con un'alta varianza nei dati saranno più precisi, fino però al raggiungimento dell'overfitting: in questi casi il modello si adatta quasi perfettamente ai dati osservati nella fase di allenamento, fallendo però nel prevedere dati non osservati precedentemente. Al contrario, tuttavia, i modelli molto semplici con poca varianza all'interno dei dati, introdurranno molto facilmente un bias, amplificandolo nella generazione dell'output.

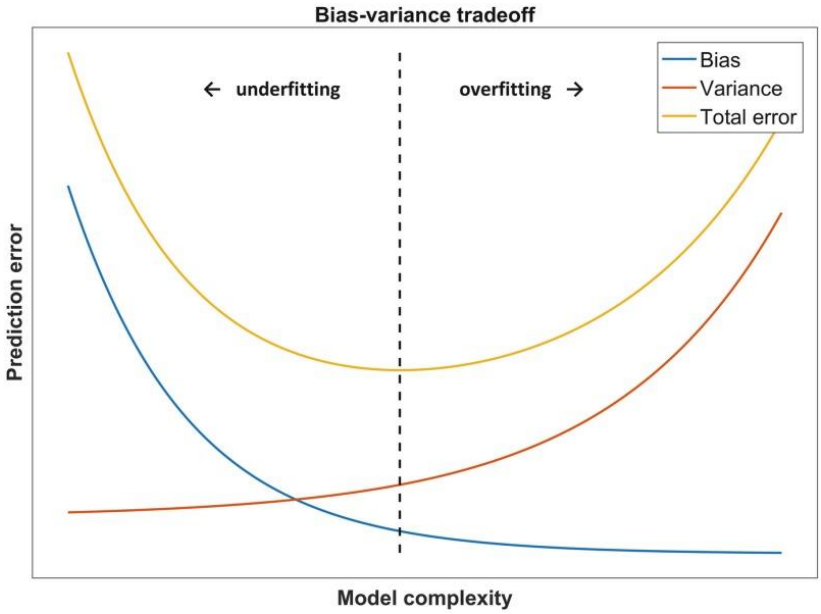

Figura 2: compromesso tra Bias e Varianza all'interno dei dati.

Fonte: ncbi.nlm.nih.gov

Diversi sistemi di riconoscimento facciale, come quelli di Microsoft e IBM, hanno avuto problemi nel riconoscimento dei volti quando si trattava di persone di colore e, soprattutto, donne. Per risolvere il problema, le aziende si sono impegnate da un punto di vista tecnico: IBM, in particolare, ha cercato di risolvere creando un dataset maggiormente inclusivo, ovvero contenente un maggior numero di dati rappresentativi di gruppi minoritari come persone di colore e donne. Per farlo sono state utilizzate milioni di immagini presenti all'interno del sito Flickr[32] e il dataset ha preso il nome di DiF (Diversity in Faces). Al fine di ottenere una migliore performance, sono state prese in considerazione le distanze tra gli elementi del volto e il genere di appartenenza delle persone presenti nei dati, ovvero uomo e donna. Questa distinzione binaria, avente lo scopo di creare una maggiore diversificazione, ha avuto però un effetto contrario, creando così un ulteriore discriminazione: tutte le persone che non appartenevano alla classificazione binaria di genere, infatti, sono state rimosse dal dataset, screditando l'esistenza della comunità trans e non binaria. Questo meccanismo

[32] Sito per il caricamento di immagini e video.

ha privilegiato l'accuratezza a discapito della *fairness*, ovvero una corretta rappresentazione della realtà in tutte le sue sfaccettature[33].

3.1.1 Le origini del bias artificiale

Il concetto di bias, osservato all'interno degli algoritmi di intelligenza artificiale, ha origini molto antiche. Già negli anni Settanta, a Londra, il dottore Geoffrey Franglen iniziò a scrivere un algoritmo per vagliare le diverse domande di ammissione degli aspiranti studenti di medicina. Al tempo Geoffrey Franglen lavorava presso la facoltà di medicina dell'ospedale St. George a cui fecero domanda di ammissione più di duemila studenti. La maggior parte di questi studenti veniva scartata nel processo di selezione iniziale, mentre, circa il 70% di coloro che passavano la prima fase, che si basava su una domanda di ammissione scritta, venivano inseriti all'interno della scuola per uno stage iniziale. Superare la fase iniziale, ovvero la domanda di ammissione scritta al fine di ottenere un secondo colloquio, risultava dunque cruciale ai fini dell'ammissione. Tuttavia, la scrematura iniziale delle domande di ammissione risultava essere la parte più impegnativa, e dispendiosa di tempo, anche per la commessione che si occupava della selezione dei candidati, di cui faceva parte anche Geoffrey Franglen. Per questo motivo, Franglen, reputandolo un lavoro adatto, decise di automatizzare la fase iniziale del processo di selezione attraverso l'utilizzo di algoritmi predittivi, scrivendo un programma che imitasse il comportamento della commissione in fase di selezione. Tuttavia, l'obiettivo di Franglen non era soltanto quello di rendere il processo di selezione più veloce ed efficiente, ma anche quello di rimuovere eventuali inconsistenze nel processo, spesso causate dal modo in cui le persone gestivano le diverse domande: l'idea portante dietro la scrittura del programma era che la macchina, contrariamente all'essere umano, non poteva essere soggetta ad influenze personali ed esterne di nessun tipo, garantendo così una valutazione totalmente equa delle domande in entrata.

Nel 1979 il programma viene testato per la prima volta: le selezioni vengono passate al vaglio sia della macchina sia della commissione per paragonarne e osservarne la

[33] J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 2018.

prestazione: alla fine del processo di selezione la macchina concordava con la commissione circa il 95% delle volte, facendo credere all'autore di avere costruito con successo un programma in grado di replicare il lavoro degli esaminatori.
Tuttavia, come già osservato precedentemente, l'autore del programma si sbagliava: nel prendere decisioni, la macchina, non è più equa degli esseri umani che la costruiscono ed allenano attraverso i loro dati.
A distanza di qualche anno, il programma veniva utilizzato come unico mezzo di selezione delle domande di ammissione, sostituendo completamente la presenza umana. Questo cambio nel processo decisionale portò a conseguenze gravi, che furono visibili soltanto dopo diversi anni di impiego: i membri della commissione, infatti, si accorsero di una mancanza di varietà etnica tra gli studenti a cui veniva concesso un posto all'interno della scuola e decisero di aprire un'indagine interna.
L'indagine portò alla scoperta di diversi pregiudizi all'interno della macchina: il sistema, infatti, sembrava giudicare gli studenti sulla base di caratteristiche non realmente rilevanti ai fini dell'ammissione, come il nome e la data di nascita, rendendo evidente alla commissione che la macchina votasse maggiormente a favore quando il nome del candidato era di origine caucasica e, allo stesso tempo, scalasse 15 punti quando il nome non risultava essere di origine europea. La macchina, dunque, rifletteva e amplificava il bias precedentemente esistente all'interno delle persone facenti parte della commissione di selezione e il programma fu accusato di essere discriminatorio verso le persone di colore e le donne[34].

I dati codificano, dunque, una serie di caratteristiche umane sotto forma di valori, in cui possono essere presenti caratteristiche che identificano motivi di discriminazione o pregiudizio. La non curanza di tali caratteristiche comporta la creazione di modelli distorti principalmente su determinate variabili piuttosto che altre, rendendo così fondamentale la comprensione delle diverse influenze causali tra le variabili e la classe di arrivo. Tuttavia, specialmente nei big data, molti dataset non sono creati con il rigore di uno studio statistico, ma sono il sottoprodotto di altre attività aventi diversi obiettivi, spesso operativi. L'intelligenza artificiale dovrebbe rappresentare un'opportunità per

[34] M. Garcia, *Racist in the machine,* 2017.

creare un futuro più equo e non un ulteriore inibitore dell'uguaglianza sociale.

3.2 Tipologie di bias all'interno dei dati

Le tipologie di bias all'interno dei dati possono essere molteplici, dal momento che, probabilmente, qualsiasi dataset ne contiene qualcuno. Tuttavia, ci sono delle tipologie di bias che vengono riscontrate spesso nei dati, portando a gravi conseguenze per gli individui e per la società. Di seguito, verranno osservate alcune tra le tipologie più frequenti nella società odierna, con un'ulteriore attenzione alla presenza del bias di ancoraggio nei dati:

-Bias razziale: Al giorno d'oggi, la discriminazione razziale verso gruppi minoritari, si può incontrare sul posto di lavoro, nell'istruzione e nella sanità, nonostante il razzismo sia stato proibito legalmente dal 1960. Il razzismo trova le sue radici nella convinzione umana che esistano diverse razze che variano da un punto di vista fisico e psicologico, nonostante il concetto di razza sia legato ad un costrutto puramente sociale, che ha portato conseguenze molto gravi in passato per diversi gruppi di individui giustificando attività come la schiavitù ed il colonialismo.

Sono numerosi i casi in cui il razzismo umano ha influenzato in qualche modo gli algoritmi di apprendimento automatico e il loro output, in particolare, uno dei casi più evidenti di pregiudizio razziale si può trovare nelle applicazioni di riconoscimento facciale: spesso, infatti, questi programmi hanno difficoltà nel riconoscere accuratamente il volto di persone di colore, in particolare donne[35]. Nella figura 3 è possibile osservare la performance delle tecnologie di riconoscimento facciale di diverse aziende.

[35] J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 2018.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Figura 3: Risultati delle diverse tecnologie di riconoscimento facciale.

Fonte: ars.electronica.art

Un ulteriore contesto, in cui l'applicazione dell'intelligenza artificiale può portare ad un output contenente una discriminazione razziale, è quello riguardante la criminalità: in paesi come l'America, infatti, è più probabile che le persone di colore vengano arrestate o incarcerate rispetto a persone bianche. Di conseguenza, le persone di colore sono presenti in larga quantità nelle foto segnaletiche utilizzate per fare delle predizioni.

-Bias di genere: Secondo gli studi effettuati dalla World Bank Group, nel 2022, il 40% dei paesi del mondo limita i diritti di proprietà delle donne: in 19 paesi, infatti, le donne non hanno pari diritti di proprietà sui beni immobili, in 43 paesi le donne vedove non hanno lo stesso diritto degli uomini nell'ereditare i beni del coniuge, 42 paesi impediscono alle figlie di ereditare il patrimonio allo stesso modo dei figli maschi e in 18 paesi il marito ha il controllo amministrativo sui beni coniugali. Risulta quindi molto più difficile per le donne avere la possibilità di prendere decisioni economiche ed avere uno stipendio adeguato.

Tuttavia, i diritti di proprietà sono un fattore chiave nello sviluppo economico di un paese e, nei paesi con una maggiore uguaglianza economica tra uomo e donna, le donne hanno un numero più elevato di proprietà rispetto agli uomini. Quando le

donne hanno la possibilità di accedere a diversi beni quanto gli uomini, infatti, le comunità prosperano, aumentando il numero e la crescita delle imprese garantendo il credito. Questo permette alle donne di investire nelle proprie famiglie, in modo da cambiare anche il futuro dei propri figli, oltre che a garantire libertà e dignità alla donna[36].

Alcune ricerche[37] hanno mostrato una correlazione positiva tra il concetto di dominanza e mascolinità: secondo luoghi comuni, infatti, gli uomini dovrebbero essere decisi, forti e assertivi, mentre le donne dovrebbero essere premurose e comprensive. La figura del leader è solitamente caratterizzata da attitudini quali la dominanza e la competitività, motivo per cui solitamente viene attribuita maggiormente agli individui di sesso maschile. In molti credono ancora che il ruolo della donna sia quello di prendersi cura dei figli e della casa, mentre quello dell'uomo di garantire una stabilità finanziaria.

Questa divisione sociale ha permesso l'introduzione di un bias di genere all'interno della forza-lavoro: è stato dimostrato che le donne a cui vengono assegnati agenti di prestito di sesso maschile, infatti, hanno solitamente tassi di interesse più elevati, importi di prestito inferiori e scadenze più vicine. Questo pregiudizio si riflette, di conseguenza, nei processi automatizzati attraverso algoritmi di apprendimento automatico: I risultati indicano che la probabilità di ottenere punteggi di credito inferiori è più elevata per quanto riguarda gli individui di sesso femminile.

Il concetto di genere, negli ultimi anni, non rientra più in una classificazione di tipo binario (uomo o donna), ma esistono diverse identità di genere, come quella transgender, tuttavia, diverse aziende utilizzano una classificazione di tipo binario quando si effettua una domanda di assunzione, non rispecchiando il fluido approccio moderno al genere[38]. Inoltre, come già osservato precedentemente, le tecnologie di riconoscimento facciale di Amazon, Microsoft e IBM fanno fatica nel classificare correttamente persone non binarie, ovvero

---

[36]World Bank Group, *The World Bank in Gneder*, 11 ottobre 2022, https://www.worldbank.org/en/topic/gender/overview

[37]S. Zaccaro, C. Kemp, P. Bader, *Leader traits and attributes*, 2004.

[38] J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 2018.

che non appartenenti né al genere femminile né a quello maschile, riportando un tasso di precisione di circa il 30% nella classificazione di uomini transgender.
Infine, nei paesi con reddito basso le donne hanno l'8% in meno di probabilità rispetto agli uomini di possedere un telefono cellulare o navigare su internet, questo comporta una mancanza di dati collezionabili riguardanti le donne.

-Bias di età: In Europe, grazie all'intervento dell'Unione Europea, è stata adottata una direttiva che prevede un eguale trattamento, nel mondo del lavoro, per tutte le differenti fasce di età. Tuttavia, questa direttiva non è sempre rispettata, dal momento che gli impiegati più anziani vengono solitamente descritti come maggiormente lenti, con minore capacità di adattamento, meno creativi e meno aperti alla formazione. Ad avere un maggior numero di pregiudizi nei confronti delle persone anziane sembrano essere proprio i giovani, contrariamente agli anziani che, avendo già sperimentato i diversi livelli di età ed essendo quindi più comprensivi, hanno atteggiamenti meno negativi e stereotipati nei confronti dei colleghi più giovani. Di conseguenza, risulta essere più probabile che siano i giovani ad ostacolare l'assunzione o la promozione di persone più anziane e non il contrario, dal momento che la maggior parte degli anziani sembra vedere i giovani come propri eguali[39].
Il bias relativo all'età dell'individuo si può trovare anche all'interno del settore finanziario: diversi studi[40] dimostrano che l'erogazione di un prestito ipotecario viene concessa più facilmente con il crescere dell'età, discriminando in questo senso le persone più giovani, poiché spesso percepite come meno stabili finanziariamente.
Per quanto riguarda, invece, il settore giudiziario e criminale, i giudici tendono a considerare gli individui tra i 18 e i 20 anni come ancora recuperabili, poiché, data la loro giovane età, sono più facile da modellare e plasmare, risultando meno pericolosi per la comunità. Al contrario, gli individui appartenenti ad una fascia di età né troppo giovane né troppo anziana, rappresentano il gruppo a cui viene

[39] Finkelstein, L. Burke, M. Raju, *Age discrimination in simulated employment contexts: An integrative analysis,* 1995.

[40] H. Black, R. Schweitzer, M. Lewis, *Discrimination in Mortgage Lending,* 1978.

solitamente inflitta una condanna più dura, a partire dall'età di 30 anni, invece, la condanna diventa meno dura con il crescere dell'età.
Il bias riguardante l'età è, dunque, presente in numerosi settori e può, di conseguenza, essere trasmesso agli algoritmi di intelligenza artificiale attraverso i dati. Inoltre, anche in questo caso, vi è una mancanza di dati collezionabili provenienti da persone anziane dal momento che gli individui di una certa età tendono ad utilizzare meno dispositivi tecnologici: anche per questo motivo, le applicazioni vengono sviluppate basandosi maggiormente sulle esigenze e preferenze dei giovani, creando così un divario generazionale sempre maggiore.

-Bias di disabilità: In diversi paesi del mondo, la disabilità è alla base di discriminazioni in diversi ambiti, soprattutto in quello lavorativo.
La disabilità, come caratteristica presente o meno in un individuo, può essere molto difficile da classificare, dal momento che ne esistono diverse tipologie che differiscono tra loro in termini di intensità e impatto sull'individuo. Inoltre, la disabilità di un individuo po' cambiare, in termini di intensità e condizioni, nel corso del tempo, rendendo ancora più complessa la sua identificazione e classificazione: anche in gruppi di dati apparentemente omogenei, infatti, sono solitamente presenti numerosi valori anomali, anche detti *outliers.* Gli outliers rappresentano valori solitamente difficili da classificare, poiché si discostano di molto dalla media dei valori presenti all'interno di un dataset: dal momento che gli algoritmi basano il loro apprendimento sull'individuazione di gruppi omogenei basati su caratteristiche simili, i dati anomali vengono spesso rimossi.
Dal momento che i dati riguardanti la disabilità includono un numero elevato di dati che si discostano dalla media del gruppo, è facile che l'algoritmo li interpreti come valori anomali, anziché classificarli come una diversa tipologia di disabilità. Contrariamente, se si crea un modello molto complesso in grado di identificare ogni tipologia di disabilità presente nei dati, è possibile che, come già osservato precedentemente, il modello vada in overfitting non riuscendo, quindi, a classificare adeguatamente nuovi dati non osservati precedentemente. Per questo motivo, si preferiscono solitamente modelli più semplici, nonostante la qualità dell'output potrebbe risultare non abbastanza elevata da rappresentare ogni individuo.
Dichiarare la propria disabilità, in ambito lavorativo, non è obbligatorio nella compilazione di domande di assunzione e si può solitamente scegliere se dichiararla

o meno. Tuttavia, è comunque possibile che le aziende si accorgano della presenza di una disabilità attraverso altre informazioni correlate ad essa come, ad esempio, la richiesta di strumenti adeguati alla compilazione di un test, che lasciano intendere la presenza di limitazioni fisiche nell'individuo.

Nonostante la privacy delle persone con disabilità sia, quindi, in qualche modo tutelata, allo stesso tempo risulta però fondamentale collezionare i dati in modo da trarre conclusioni statistiche: secondo diversi studi, infatti, le persone che dichiarano la propria disabilità durante il percorso di assunzione hanno circa il 30% in meno di possibilità di assunzione rispetto a chi non lo fa, e uno stipendio inferiore, anche nel caso in cui la loro disabilità non grava sulla produttività lavorativa[41].

Un'ulteriore discriminazione avviene nell'utilizzo di programmi per il riconoscimento vocale ed il riconoscimento facciale: questa tipologia di programmi, infatti, può non funzionare correttamente nel momento in cui il modo di parlare di una persona, il suo aspetto o il suo comportamento si discostano dalla norma, portando alla necessità di creare modelli appositi.

-Bias di ancoraggio: Il bias di ancoraggio è presente soprattutto all'interno di processi decisionali, nonostante passi spesso inosservato. Secondo il bias di ancoraggio, infatti, spesso le persone, nel prendere decisioni, risultano ancorate ad elementi osservati precedentemente, che influiscono in qualche modo sulla loro scelta finale. Di conseguenza, il bias di ancoraggio si può trovare soprattutto nei dati che rappresentano attività in cui bisogna prendere o decisioni continue, come ad esempio nella revisione di domande di ammissione o di prestito, o decisioni basate, anche involontariamente, su dati osservati in precedenza. Gli utenti, infatti, possono essere ancorati dalle loro stesse decisioni prese precedentemente, come avviene nei processi decisionali sequenziali.

[41] M. Ameri, L. Schur, D. Kruse, *The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior*, 2017.

Tra le tecniche di mitigazione si possono, tuttavia, mettere le persone al corrente del proprio livello di ancoraggio durante il processo decisionale, nonostante questo possa comportare poi il subentro di altri bias cognitivi nell'individuo.

La presenza di questi bias all'interno dei dati può dunque comportare conseguenze dannose per gli individui. In particolare, nel campo del lavoro, in quello giudiziario, in quello sanitario e in quello finanziario, dal momento che, in questi ambiti, un piccolo errore può drasticamente cambiare la vita di una persona. Non sono infatti mancate occasioni in cui la presenza di bias all'interno di algoritmi e dati ha portato a conseguenze dannose nella vita reale.
Nel capitolo a seguire verranno osservati i casi più famosi degli ultimi anni, in cui un'errata gestione dei dati ha creato problematiche reali.

3.3 Principali casi di discriminazione

Nel corso degli ultimi anni, ci sono stati diversi casi in cui la presenza di bias, all'interno di algoritmi o dati, ha portato a conseguenze dannose per gruppi di individui. Tra i più famosi si ricordano COMPAS, PredPol e gli algoritmi di Amazon e Google foto.

COMPAS è un algoritmo sviluppato dall'azienda Northpointe negli anni Novanta, utilizzato nei tribunali statunitensi come strumento in grado di prevedere il rischio di recidività di un individuo. Il software veniva utilizzato dai giudici all'interno dei tribunali dopo che l'individuo veniva arrestato. Il giudice poneva all'individuo una serie di domande e le risposte venivano date manualmente in input al software: molte di queste domande venivano compilate automaticamente dal programma, dal momento che, inserendo il documento d'identità dell'individuo, era possibile estrapolare automaticamente delle informazioni come, ad esempio, la fedina penale, l'età, il sesso. Una volta inserite tutte le risposte necessarie, il software analizzava le informazioni in suo possesso per stabilire una probabilità di recidività compresa in un intervallo tra 1 e 10: a seconda della probabilità stimata, il giudice assegnava all'individuo un percorso di riabilitazione adeguato. COMPAS, dunque, era un valido strumento di supporto sia nel calcolo della recidività sia nella scelta del percorso di riabilitazione.

Nel 2016, l'algoritmo è stato accusato di essere discriminatorio verso le persone di colore, sollevando diverse polemiche sull'impatto negativo che l'apprendimento automatico può avere sulla società. ProPublica, redazione indipendente di giornalismo investigativo, analizzò i dati delle persone classificate utilizzando COMPAS scoprendo che, nonostante la razza di appartenenza non fosse tra le domande nella compilazione iniziale, alle persone di colore veniva assegnata una percentuale più alta di rischio di recidività rispetto alle persone bianche.
I dati forniti in input al programma riguardavano principalmente la fedina penale dell'individuo ed erano dati generati dall'essere umano: di conseguenza, quando era già presente una discriminazione verso le persone di colore nelle pratiche precedenti all'utilizzo del programma, questa si rifletteva anche nell'algoritmo amplificandola. COMPAS rappresentava, dunque, una serie di relazioni tra il codice e l'ambiente in cui veniva utilizzato.

L'algoritmo PredPol è stato sviluppato in America con l'obiettivo di ottenere una predizione più equa delle aree in cui è più probabile si sviluppino dei crimini, utilizzando algoritmi di apprendimento automatico. L'algoritmo, infatti, utilizza l'intelligenza artificiale per predire le zone in cui verranno commessi crimini, basandosi su dati precedentemente raccolti dalla polizia, come, ad esempio, il numero di arresti e chiamate effettuati nella zona.
Tra il 2018 e il 2021 numerosi statunitensi sono stati soggetti a controlli e pattugliamenti stabiliti dal software PredPol: i residenti dei quartieri in cui il software suggeriva controlli con meno frequenza risultavano essere per la maggioranza bianchi e benestanti, contrariamente alle zone abitate maggiormente da persone di colore e latini che, invece, venivano segnalati senza sosta dal sistema per continui controlli ogni giorno, anche più volte al giorno, in più luoghi dello stesso quartiere, portando così a migliaia di previsioni anche in momenti in cui non era necessario.

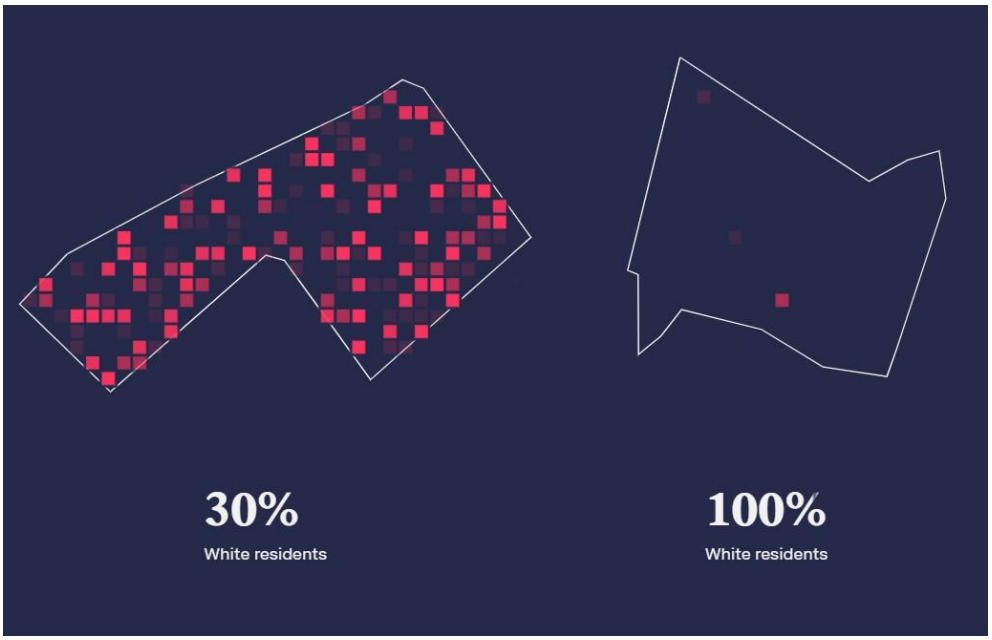

Figura 4: Previsioni PredPol di due diversi quartieri di Los Angeles.
Fonte: themarkup.org

I quartieri del Michigan, in cui l'algoritmo prevedeva maggiori pattugliamenti, avevano un numero di residenti di colore pari a sette volte sopra la media della città; in Alabama, dove circa la metà dei residenti consisteva in persone di colore, vi era un significante numero inferiore di segnalazioni, da parte del sistema, nelle aree prevalentemente abitate da individui bianchi; a Losa Angeles, dove il sistema sembrava segnalare aree abitate maggiormente da persone bianche, il pattugliamento si concentrava in zone con popolazioni totalmente Latine; a Chicago, i quartieri segnalati meno frequentemente dalla macchina erano solitamente abitati da individui molto ricchi, con un reddito superiore alla media, mentre i quartieri popolati da persone con un reddito basso ricevevano più del doppio delle segnalazioni e gran parte di loro era di origine Latina[42]. Nel complesso, quindi, PredPol prevedeva una maggiore possibilità di crimine nelle aree dove i residenti erano maggiormente o neri o Latini o poveri, creando una disparità basata, non soltanto sull'etnia, ma anche sul reddito.

[42] A. Sankin, D. Mehrotra, S. Mattu, D. Cameron, A. Gilbertson, D. Lempres, J. Lash, *Crime Prediction Software Promised to Be Free of Biases*, dicembre 2021, GIZMODO, https://gizmodo.com/crime-prediction-software-promised-to-be-free-of-biases-1848138977

Dal 2014 Amazon ha sviluppato algoritmi con lo scopo selezionare automaticamente le domande di assunzione dei migliori candidati. Nel 2015, l'algoritmo si occupava di analizzare i curriculum dei candidati interessati ad entrare in Amazon, per selezionare quali di loro sarebbero passati alla fase di selezione successiva, assegnando a ciascun candidato un punteggio da 1 a 5, proprio come accade nelle recensioni dei loro prodotti.

Nel corso dell'anno, l'azienda si rese conto che la macchina, nell'assegnare i punteggi per lavori maggiormente tecnici, non era neutrale dal punto di vista del genere, favorendo maggiormente candidati di genere maschile. Questo accadeva perché il modello utilizzato era stato allenato su dati provenienti da curriculum raccolti dall'azienda nel corso di dieci anni e la maggior parte di questi curriculum, soprattutto nel settore tecnologico, erano di individui maschili. Di conseguenza, l'algoritmo rifletteva la dominanza maschile nel settore. Il sistema aveva, quindi, imparato che l'azienda preferiva gli uomini alle donne, penalizzando i curriculum che avevano al loro interno parole relative alle donne o nomi di università con studenti prevalentemente donne.

Il modello era stato sviluppato nella sede dell'azienda a Edimburgo, con l'obiettivo di scannerizzare velocemente il web in modo da trovare i candidati migliori per determinate posizioni lavorative. Il team responsabile allenò il modello a riconoscere più di 50000 termini presenti nei curriculum dei precedenti candidati: tuttavia, in questo modo, l'algoritmo imparò a dare meno rilevanza ai termini riguardanti abilità pratiche comuni nel campo tecnologico, come, ad esempio, la conoscenza di diversi linguaggi di programmazione, favorendo, invece, i candidati che si descrivevano con termini più comunemente utilizzati, all'interno del curriculum, da uomini ingegneri.

Amazon tentò di risolvere il problema rendendo la macchina neutrale davanti determinati termini[43].

---

[43] J. Dustin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Nel 2015, nell'aggiornamento di Google Photos[44] è stata inserita una nuova funzione che permette al programma di etichettare le foto automaticamente a seconda di ciò che viene mostrato nell'immagine. Questa nuova funzione è stata realizzata attraverso l'utilizzo di una rete neurale convoluzionale[45] allenata in modo supervisionato su milioni di immagini. Tuttavia, l'algoritmo di Google si è mostrato razzista nell'etichettare la foto in cui appariva un ragazzo di colore e un suo amico, etichettandoli come gorilla. Ad accorgersi del bias presente è stato il ragazzo presente nella foto, che ha richiamato l'attenzione dell'azienda attraverso un tweet.
Google si dichiarò dispiaciuto dell'errore promettendo di correggerlo, tuttavia, nei due anni seguenti Google aggirò semplicemente il problema, rimuovendo dalle parole date in input all'algoritmo quelle relative alle scimmie[46].
Quello di Google Photos però non fu l'unico caso di razzismo presente negli algoritmi dell'azienda: nel 2020, infatti, Google Vision Cloud, strumento di computer vision, etichettava il termometro frontale come "pistola" nelle immagini in cui questo veniva sorretto da persone di colore, e come "termometro" o "strumento elettronico" nelle immagini in cui a sorreggerlo era una persona bianca[47].

3.4 Tecniche di mitigazione

Come osservato precedentemente, la presenza di bias all'interno dei dati può influire significativamente sulla vita delle persone, motivo per il quale individuarli e, se necessario, correggerli, diventa un'attività fondamentale nell' utilizzo di un modello. Esistono molteplici approcci per eliminare il bias, nonostante nessuno di questi sia infallibile: si può, ad esempio, cercare di collezionare i dati in un modo che garantisca l'assenza di bias o scrivere algoritmi in grado di minimizzarlo.

---

[44] Servizio di condivisione e archiviazione immagini.

[45] Tipologia di rete neurale utilizzata maggiormente per l'analisi di immagini.

[46] T. Simonite, *When It Comes to Gorillas, Google Photos Remains Blind,* 2018, WIRED, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[47] N. Kayser-Bril, *Google apologizes after its Vision AI produced racist results*, 2020, ALGORITHM WATCH, https://algorithmwatch.org/en/google-vision-racism/

Dal momento che nessun set di dati può rappresentare realmente l'intero universo di opzioni, è importante stabilire in anticipo l'utilizzo che se ne vuole fare e il pubblico di arrivo, adattando i dati di addestramento all'obiettivo desiderato. Identificare i bias non è semplice, come osservato nei capitoli precedenti, in alcuni casi ci si accorge del problema soltanto a posteriori, quando l'utilizzo dell'intelligenza artificiale ha già influito dannosamente nella vita di determinati gruppi. I bias possono essere individuati, ed eventualmente mitigati, prima della costruzione e dell'utilizzo del modello, in modo da creare un modello che non ne contenga (pre- processing), durante l'allenamento del modello (in-processing), oppure a posteriori, ovvero sulle predizioni fatte dall'algoritmo.

3.4.1 Pre-processing

Si possono utilizzare diversi approcci per individuare i bias presenti nei dati, o negli algoritmi, prima che questi abbiano un impatto nel mondo reale: si può, ad esempio, controllare se gruppi di persone, che potrebbero essere facilmente discriminati dall'automatizzazione, siano ben rappresentati all'interno del dataset, ovvero se il modello utilizzato riesce ad apprendere, in modo appropriato, le caratteristiche relative a quei gruppi. Inoltre, comparare la qualità dei dati relativa a determinati gruppi, solitamente a rischio di discriminazione, con quella dell'intera popolazione di dati può aiutare ad evitare la formazione di bias: nello sviluppo di un sistema di riconoscimento facciale da utilizzare all'interno degli aeroporti per controlli di sicurezza, ad esempio, il dataset utilizzato per addestrare il modello deve rappresentare i gruppi più a rischio in modo che il modello sia in grado di apprenderne informazioni sufficienti per un adeguato riconoscimento, utilizzando un ampio numero di immagini. In caso contrario, il sistema di riconoscimento facciale funzionerà in modo meno accurato sul gruppo a rischio.

Questa tipologia di approccio si concentra principalmente sui dati, con l'obiettivo di creare un dataset che sia il più possibile bilanciato nel rappresentare i gruppi presenti nella popolazione, dal momento che meno discriminatori sono i dati con cui viene allenato l'algoritmo (quindi del training set), meno sarà discriminatorio il modello in cui vengono inseriti.

3.4.2 In-processing

Nell'approccio in-processing il bias viene corretto durante l'allenamento del modello, riformulando il problema di classificazione attraverso l'utilizzo di una funzione obiettivo. Una funzione obiettivo è una funzione utilizzata per massimizzare o minimizzare qualcosa, in modo da ridurre la differenza tra i valori attesi[48] e quelli reali. Solitamente questa funzione viene utilizzata per massimizzare i profitti o minimizzare le perdite, inserendo un insieme di vincoli e basandosi sulla relazione tra una o più variabili influenti sulla classe di arrivo. Quando si utilizzano i modelli a *scatola bianca*[49] , un possibile approccio è quello di alterare il modello internamente: correggendo, ad esempio, le probabilità in un modello probabilistico come quello del Naive Bayes, o correggendo la classe di appartenenza dei dati nelle foglie di un Decision Tree.

Solitamente queste tecniche fanno riferimento ad un apprendimento supervisionato, ma è possibile utilizzare anche approcci non supervisionati come, ad esempio, la PCA, ovvero una tecnica di riduzione della dimensione dei dati in uno spazio minore, forzando un'equa ricostruzione nello spazio ridotto sia dei gruppi a rischio che i gruppi non a rischio.

3.4.3 Post-processing

L'approccio Post-processing, invece, consiste nell'alterazione delle predizioni fatte dall'algoritmo. In questi casi, solitamente, si fa riferimento ad algoritmi a scatola nera: un possibile approccio consiste nell'alterazione dei punti (o dati) posizionati attorno al decision boundary dell'algoritmo (figura 5), ovvero il confine che determina la separazione tra le classi: in breve, più un punto si trova vicino a questo confine, minore sarà la sua confidence, ovvero la sicurezza, da parte dell'algoritmo, di appartenenza di quel punto alla classe in cui è stato classificato.

---

[48] Valore medio previsto dopo un elevato numero di prove.

[49] Modelli non a scatola nera di cui, quindi, è facile interpretare la logica utilizzata nel prendere decisioni, come ad esempio negli alberi decisionali (o Decision Tree).

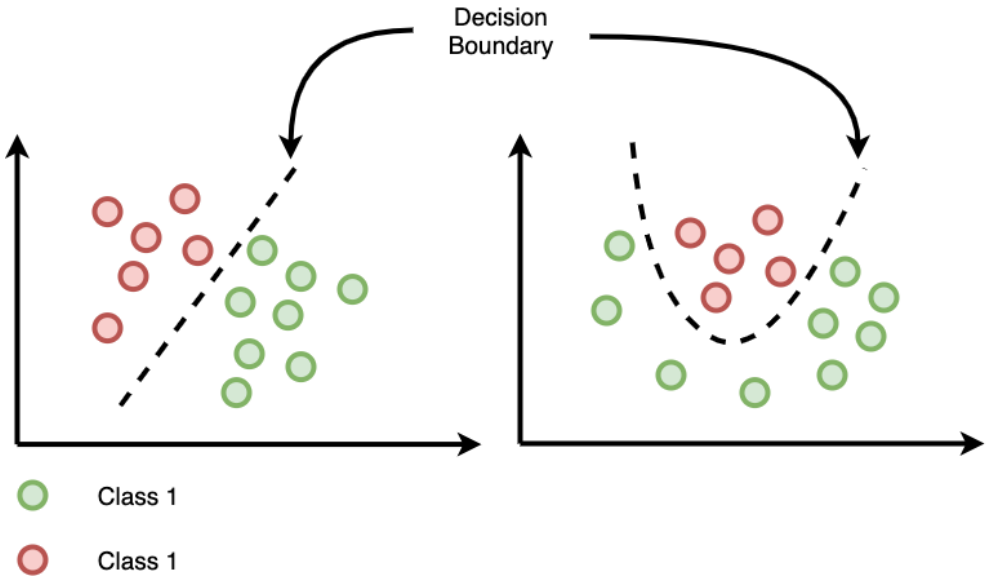

Figura 5: Decision Boundary.

Fonte: towardsdatascience.com

In alternativa, un'altra possibile soluzione, è quello di sostituire la black box dell'algoritmo con un modello più facilmente interpretabile, oppure utilizzare degli algoritmi in grado di spiegare cosa accade dentro la scatola nera, per poi modificare il necessario di conseguenza.

Quindi, mentre la maggior parte degli approcci di mitigazione del bias si concentrano sull'utilizzo di modelli supervisionati o facilmente interpretabili, gli approcci di tipo post-processing sono rilevanti quando si utilizzano algoritmi di tipo non supervisionato o a scatola nera.

3.4.4 Implicazioni legali:

L'utilizzo degli approcci precedentemente osservati al fine di mitigare il bias all'interno della macchina, solleva, tuttavia, diverse questioni di tipo legale.

Queste tecniche, infatti, prevedono la modifica dei dati o l'alterazione del modello utilizzato senza che vi sia un'opportuna regolamentazione sul fatto che questi approcci possano essere o meno considerati leciti. Non esiste, infatti, alcuna disposizione legale che si occupi di regolamentare il modo in cui i dati, in questi casi, vengono raccolti, selezionati o modificati, tuttavia, la legittimazione potrebbe essere ottenuta attraverso un consenso informato.

Un' ulteriore problema, che può crearsi durante l'attenuazione di bias, è quando la sua mitigazione coinvolge dati sensibili, come, ad esempio, l'etnia. In proposito si esprime l'articolo 9 del GDPR, che consente il trattamento di questi dati per motivi di interesse pubblico affermando quanto segue: *"il trattamento è necessario per motivi di interesse pubblico rilevante sulla base del diritto dell'Unione o degli Stati membri, che deve essere proporzionato alla finalità perseguita, rispettare l'essenza del diritto alla protezione dei dati e prevedere misure appropriate e specifiche per tutelare i diritti fondamentali e gli interessi dell'interessato;"*.

Le preoccupazioni legali sono presenti anche in caso di modifica del modello: contrariamente a quanto succede quando si modificano i dati, quando si modifica il modello la legge sulla protezione dei dati non ha valenza, poiché il modello non contiene dati personali.

3.5 Aprire la black box

Gli algoritmi di intelligenza artificiali, sviluppati nel campo del machine learning, sono ormai utilizzati ovunque, anche nei settori più rilevanti per la vita delle persone, come gli istituti governativi e la medicina. Nell'apprendere i pattern presenti nei dati, come precedentemente osservato, gli algoritmi assorbono anche il bias contenuto in essi, amplificandolo.

La difficoltà sta nel comprendere quando un algoritmo possiede o meno un bias al suo interno, specialmente quando si utilizzano algoritmi a scatola nera, di cui neanche i più esperti spesso sanno stabilirne il suo livello di correttezza. Quando si utilizzano algoritmi più complessi, infatti, è quasi impossibile stabilirne la logica, tanto che neanche gli stessi creatori ne conoscono esattamente il funzionamento. Il problema della scatola nera impedisce, quindi, di capire come l'algoritmo arriva alla conclusione di output e, se non risolto, può propagare diverse tipologie di discriminazione che la società sta cercando di lasciarsi alle spalle da diversi anni.

Le norme che regolano l'utilizzo di questi algoritmi non sono presenti in tutti i campi di applicazione: infatti, mentre alle banche è vietato utilizzarli nella valutazione di richieste di finanziamento da parte della clientela, nelle aree riguardanti, ad esempio, i servizi pubblicitari o nel sistema giuridico non vi è la presenza di questo divieto, rendendo difficile comprendere i motivi presenti dietro a determinate scelte.

La mancanza di accesso alla scatola nera impedisce, inoltre, di individuare facilmente un possibile bias presente nella macchina. In proposito di ciò, l'Unione Europea prevede che gli individui non siano soggetti a decisioni prese soltanto da algoritmi di apprendimento automatico, ma che vi sia anche un intervento umano, soprattutto quando si fa riferimento a situazioni che possono influire seriamente sulla vita delle persone. Inoltre, è in aumento la richiesta di sviluppo di modelli a scatola bianca sempre migliori, in modo da superare l'ostacolo dell'utilizzo riguardante i modelli a scatola nera, e permettere a diverse organizzazioni di progredire. Purtroppo, però, è difficile per gli algoritmi meno complessi raggiungere un alto livello di *accuracy*[50] quando si ha un numero molto elevato di features e, di conseguenza, ci si trova in una situazione di stallo in cui si deve scegliere se utilizzare un modello con una minore accuratezza ma facilmente interpretabile, oppure un modello con una maggiore accuratezza che non permette però di comprendere il suo funzionamento interno. L'imposizione di regole troppo restrittive rischia quindi di danneggiare la competitività del sistema finanziario: secondo Banca Intesa Sanpaolo, infatti, attraverso l'utilizzo di tecniche ed elaborazioni innovative si potrebbe ottenere un'accuratezza del 10% in più rispetto alla media[51].

Negli ultimi anni si è assistito allo sviluppo di una nuova branca del machine learning, che prende il nome di *explainability*, e consiste nello sviluppo di algoritmi in grado di sostituire la scatola nera del modello o di "sbirciare" al suo interno, in modo da osservarne la logica. Lo sviluppo di questi modelli permetterebbe, quindi, non solo di utilizzare gli algoritmi a scatola nera superando la questione etica, ma anche di individuare più facilmente la presenza di possibili bias, osservando quali variabili hanno avuto una maggiore rilevanza sull'output.

---

[50] Misura di valutazione del modello che permette di quantificare quanto le predizioni si avvicinino al valore reale.

[51] D. Aliperto, *Prestiti e machine learning, sul cammino i nodi privacy e black box,* 2022, CORCOM, https://www.corrierecomunicazioni.it/finance/prestiti-digitali-alla-prova-machine-learning-ma-vanno-sciolti-i-nodi-privacy-e-black-box/

4 Individuazione dell'anchoring bias all'interno dei processi decisionali

Il capitolo si concentra sull'individuazione del bias di ancoraggio all'interno dei processi decisionali, attraverso l'utilizzo di due dataset differenti. Il primo dataset preso in considerazione raccoglie i dati delle domande di ammissione per un Master da parte degli studenti, mentre il secondo dataset consiste in una raccolta di votazioni di film da parte degli utenti. In questi casi il bias è strettamente legato all'ordine in cui vengono visionati gli elementi da valutare, l'ancoraggio, infatti, si crea in base alla sequenza in cui gli esaminatori visionano le domande di ammissione o i film da recensire: quando viene analizzata una domanda o recensito un film borderline[52], l'esaminatore si dimostra più incline a dare una valutazione positiva se le istanze subito precedenti sono state valutate negativamente. Al contrario, se le istanze precedenti a quella borderline sono state valutate positivamente, l'esaminatore si dimostra più incline a rifiutare il candidato o a recensire il film negativamente. Di conseguenza, maggiore è la distanza dall'ultima istanza positiva, maggiore è la probabilità che l'istanza borderline venga valutata positivamente e viceversa. Una volta individuato il bias, se presente all'interno dei dati, si procede con l'applicazione di un algoritmo di explainability, in grado di sbirciare all'interno della scatola nera. L'obiettivo è quindi quello di individuare i possibili dati ancorati e predire l'ancoraggio su dati "nuovi", per poi osservare le features che hanno portato a questa decisione.

4.1 Studi di riferimento e obiettivi

Questa ricerca prende spunto dal paper *AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks*[53] in cui gli autori individuano e mitigano il bias di ancoraggio. Tuttavia, in questa tesi, ci si concentrerà soltanto in modo teorico su una possibile mitigazione del bias, esplorando invece la parte riguardante il problema etico, non analizzata nel paper di riferimento.

[52] Caso limite.

[53] J. Echterhoff, M.Yarmand, J. McAuley, 2022.

L'articolo menzionato si concentra sull'individuazione del bias di ancoraggio su due tipologie di dataset differenti: il primo raccoglie i dati delle domande di ammissione da parte degli studenti ad un college americano, mentre il secondo consiste in un esperimento in tempo reale di recensione di diversi prodotti da parte delle persone. Nell'esperimento riguardante la recensione di diversi prodotti, lo stato di ancoraggio è dato dall'influenza che le valutazioni precedenti hanno sull'utente.

Il bias di ancoraggio prevede, come già osservato precedentemente, che le persone si ancorino ad un primo dato iniziale durante i processi decisionali, regolando involontariamente su quel primo dato tutte le scelte a seguire.

Nel paper, al fine di individuare il bias di ancoraggio, gli autori utilizzano inizialmente un Support Vector Machine a cui vengono dati in input i dati: la confidence dell'algoritmo diminuisce all'aumentare della distanza tra l'ultima istanza positiva o negativa e quella borderline: maggiore è la distanza, minore è la confidence dell'algoritmo nell'analizzare le istanza successive, dimostrando in un certo senso l'ancoraggio che avviene nell'esaminatore e l'importanza dell'ordine in cui vengono visionate le istanze. Le predizioni date dall'algoritmo vengono poi inserite in input all'interno di una rete neurale LSTM[54] che ritorna sottoforma di probabilità lo stato di ancoraggio. Questa probabilità viene successivamente utilizzata come parametro di mitigazione del bias all'interno di un modello di reinforcement learning[55] che cambierà l'ordine in cui vengono presentate le istanze in modo che il bias venga minimizzato.

Quando si utilizzano tecniche di questo tipo si creano diverse problematiche etiche relative alla loro applicazione nel mondo reale: l'utilizzo di algoritmi a scatola nera in contesti delicati in grado di cambiare profondamente la vita delle persone, come l'ammissione o meno ad un college, non è ben visto e difficilmente consentito, dal momento che questa tipologia di algoritmi non permette di conoscere a fondo la logica che c'è dietro all'output fornito. Di conseguenza, in questa tesi, non ci si occuperà della mitigazione del bias una volta individuato, ma si tenterà di aggirare il problema relativo alla scatola nera, applicando un algoritmo di explainability in grado di spiegare la previsione dello stato di ancoraggio data dall'algoritmo.

---

[54] Long short-term memory.

[55] Tecnica di apprendimento automatico in grado di scegliere le azioni da compiere per il conseguimento di determinati obiettivi.

4.2 Primo dataset: domande di ammissione

Il primo dataset[56] utilizzato è stato reperito attraverso la piattaforma Kaggle e, come già precedentemente accennato, è un dataset contenente dati relativi alla valutazione di domande di ammissione per un Master. Tuttavia, è necessario sottolineare che questo dataset consiste in un primo dataset di *esplorazione* dal momento che è stato, in parte, creato artificialmente.

4.2.1 Comprensione dei dati

Il dataset prende il nome di *Graduate Admission* ed è composto 400 righe e 9 colonne.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Figura 6: visualizzazione delle prime 5 righe del dataset.

Come visibile nella figura 6, il dataset contiene le seguenti features:

-GRE Score: punteggio ottenuto dal candidato nell'esecuzione di un test di comprensione del testo, scrittura e matematica.

-TOEFL Score: punteggio ottenuto dal candidato nell'esecuzione di un test di valutazione della lingua inglese.

-University Rating: punteggio da 1 a 5 assegnato dall'università.

-SOP: valutazione della dichiarazione di intenti scritta dal candidato.

-LOR: valutazione della lettera di raccomandazione del candidato.

[56] https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

-CGPA: misura del rendimento scolastico complessivo dello studente.

-Research: assume il valore 1 se il candidato ha esperienza nella ricerca, altrimenti assume il valore 0.

-Chance of Admit: probabilità di ammissione del candidato in un range tra 0 e 1.

Il dataset non contiene, inoltre, valori mancanti: nelle features di un dataset, infatti, è possibile che alcuni valori siano assenti. I motivi per cui ciò accade possono essere diversi come, ad esempio, la mancanza di un' informazione poiché non fornita dal partecipante o una mancata registrazione del valore a causa del malfunzionamento dello strumento utilizzato per raccogliere uno specifico tipo di dato.
Per avere una migliore comprensione di come le features si influenzano tra loro, è stata osservata la loro correlazione attraverso una heatmap.

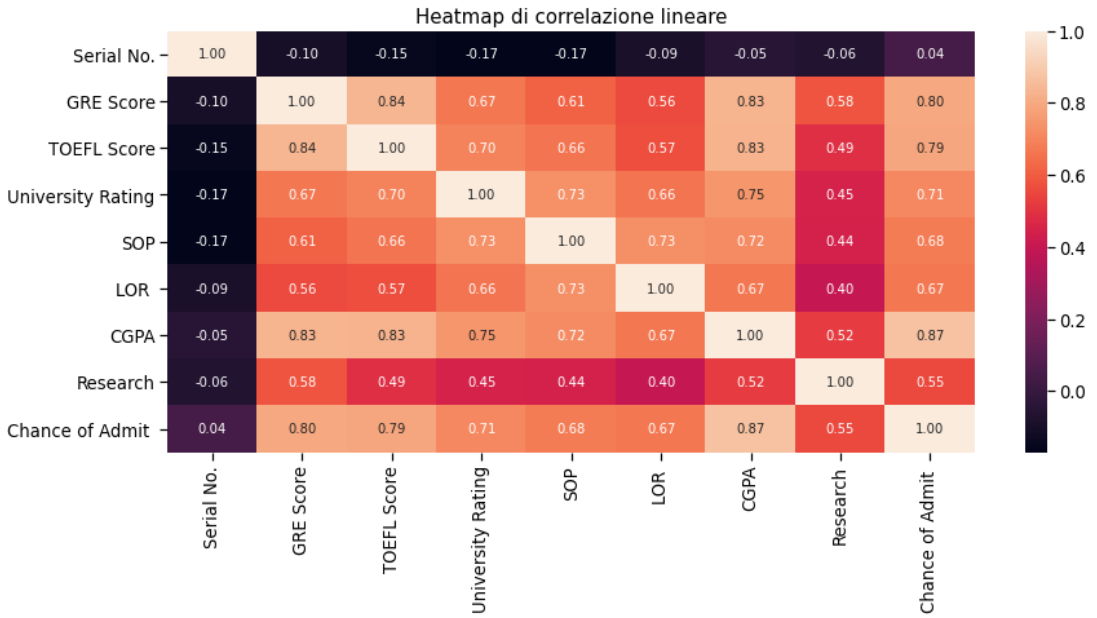

Figura 7: heatmap di correlazione lineare tra le features presenti nel dataset.

Il range di valori che la correlazione può assumere va da 0 a 1, dove il valore 0 indica una mancanza di correlazione, mentre il valore 1 indica una forte correlazione tra le variabili. La mappa mostra una forte correlazione soprattutto tra i dati presenti in *Chance of Admit* e quelli presenti in *CGPA* e *GRESCORE*, assumendo un valore di correlazione pari a 0.87 nel primo caso e 0.80 nel secondo.

4.2.2 Preparazione dei dati

Al fine di rilevare il possibile ancoraggio presente nei dati è necessario lavorare sul dataset: come scritto precedentemente questo dataset è utilizzato come dataset di esplorazione, poiché è necessario effettuare modifiche ed implementarlo al fine di raggiungere l'obiettivo desiderato.

Innanzitutto, il nome della colonna *Chance of Admit* è stato modificato in *Admission_rate* in modo da avere una migliore comprensione delle feature che verranno aggiunte. Successivamente è stata creata ed aggiunta al dataset la colonna *Admission_state*, contenente lo stato della domanda di ammissione. Il valore delle istanze in *Admission_state* dipende dal valore presente in *Admission_rate*; se compreso tra 0.66 e 0.75 (inclusi) l'*Admission_state* assume il valore *border* (borderline), se superiore a 0.75 assume il valore *in* (ammesso) e, infine, se inferiore a 0.66 assume il valore *out* (non ammesso). Attraverso questo procedimento è possibile visionare quando una domanda può essere o meno considerata borderline e, quindi, possibile vittima di ancoraggio da parte dell'esaminatore. Alla fine del procedimento risultano essere presenti 172 ammessi, 122 non ammessi e 106 candidati borderline.

Tuttavia, aggiungere la colonna *Admission_state* non è sufficiente al fine di individuare l'ancoraggio, ma è necessario creare ed aggiungere un'ulteriore colonna contenente una *ground truth*, ovvero una verità di base, o esito della domanda, a cui fare riferimento per osservare l'esito delle domande, precedenti a quella presa in considerazione, visionate dall'esaminatore. Anche per creare la colonna *Ground_truth*, che mira a simulare la decisione finale presa da un comitato, si utilizzano i valori presenti in *Admission_rate*: questa volta la nuova colonna assume un valore uguale a 1 (ammesso) se il valore assunto dall'istanza in *Admission_rate* è superiore o uguale a 0.68, altrimenti assume il valore 0 (non ammesso).

	Serial_No.	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR_	CGPA	Research	Admission_rate	Admission_state	Ground_truth
0	1	337	118	4	4.5	4.5	9.65	1	0.92	in	1.0
1	2	324	107	4	4.0	4.5	8.87	1	0.76	in	1.0
2	3	316	104	3	3.0	3.5	8.00	1	0.72	border	1.0
3	4	322	110	3	3.5	2.5	8.67	1	0.80	in	1.0
4	5	314	103	2	2.0	3.0	8.21	0	0.65	out	0.0

Figura 8: visualizzazione delle prime 5 righe del dataset post-modifiche.

Una volta applicate le modifiche, il dataset risulta formato da 11 colonne.

Un ulteriore step nella fase di preparazione dei dati è quella di suddivisione del dataset in training set e test set: prima di individuare l'ancoraggio all'interno del dataset, infatti, è stato utilizzato un algoritmo di intelligenza artificiale al fine di osservarne la performance nella predizione dell'*Admission_state.* La divisione in training e test set è stata realizzata con una proporzione 70:30, dove il 70% dei dati viene utilizzato per il training e il 30% per il test, impostando i valori assunti dalla colonna *Admission_state* come classe di arrivo.

Infine, è stato bilanciato il dataset dal momento che le classi risultavano non adeguatamente bilanciate. Uno sbilancio nella proporzione delle classi di arrivo può creare una rappresentazione non adeguata dei dati da parte dell'algoritmo utilizzato, che potrebbe non essere in grado di predire le istanze con una classe di arrivo minoritaria, eliminandola quindi dalla predizione. Questo comportamento potrebbe risultare in un'accuratezza dell'algoritmo molto alta, dal momento che la maggior parte delle istanze di classe non minoritaria viene predetta correttamente; tuttavia, il valore non sarebbe veritiero dal momento che l'algoritmo non rappresenterebbe in alcun modo la classe minoritaria. Per questo motivo esistono delle metodologie in grado di bilanciare le classi. In questo caso è stato utilizzato il metodo SMOTE. Questo metodo è una tecnica di *oversampling* [57]e consiste nell'individuazione di punti vicini appartenenti alla classe minoritaria, tra cui generare dei punti sintetici appartenenti alla stessa classe. Dopo la sua applicazione, le classi *border, in* e *out* presenti in *Admission_state* contano 127 punti ciascuna.

4.2.3 Support Vector Machine

Questo passaggio non è fondamentale per l'individuazione dell'ancoraggio, ma è utile al fine di mostrare la possibilità di automatizzazione della predizione dello stato di ammissione ed individuare in seguito l'ancoraggio sui risultati predetti dalla macchina. L'algoritmo utilizzato per la predizione prende il nome di Support Vector Machine: questo algoritmo crea un decision boundary, ovvero un confine in grado di

[57] Tecnica di bilanciamento dei dati che consiste nel portare la classe minoritaria al livello di quella maggioritaria.

separare, nel miglior modo possibile, le classi, massimizzando i margini di un iperpiano. Per stabilire i margini del decision boundary, l'algoritmo utilizza un subset di dati d'esempio che prendono il nome di support vectors.

Al fine di trovare i valori ottimali che permettono all'algoritmo di suddividere nel miglior modo i dati in input, è stata effettuata una ricerca dei migliori parametri, attraverso l'utilizzo di una gridsearch[58], così da controllare il processo di addestramento.

I parametri migliori sono risultati essere i seguenti: *C=100,gamma=1, kernel="linear", random_state=42.* Nella figura 9 è possibile vedere i risultati della classificazione dell'algoritmo attraverso l'utilizzo di una confusion matrix.

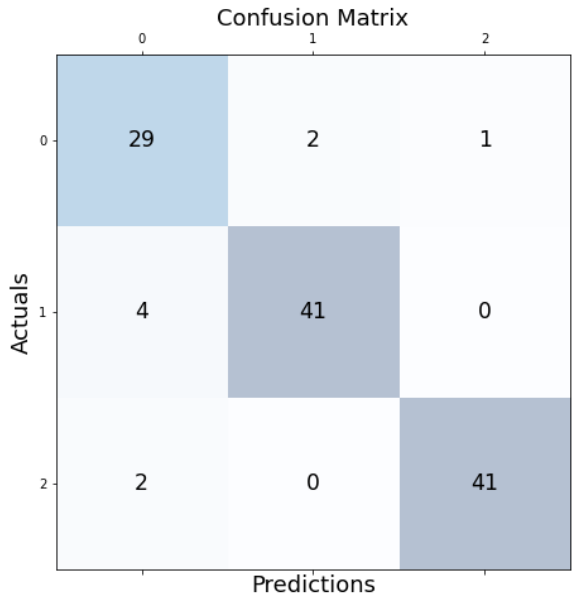

Figura 9: confusion matrix Support Vector Machine.

Nella confusion matrix il valore 0 sta per *border*, mentre il valore 1 sta per *in* e il valore 2 sta per *out*, l'accuracy della performance è di 0.925. Il classificatore performa quindi in maniera discreta sui dati in input, miss classificando tre volte la classe zero, classificandola due volte come *in* e una volta come *out*, e miss classificando quattro volte la classe uno classificandola erroneamente quattro volte come *border*. Infine, la classe 2, è stata miss classificata due volte come *border*. Attraverso la confusion

[58] Metodo utilizzato nel campo del machine learning che permette di testare diversi parametri per poi trovare quelli che meglio rappresentano i dati in input.

matrix, quindi, è possibile vedere quanti punti sono stati classificati correttamente nella classe di appartenenza e quanti sono stati classificati erroneamente come una classe diversa da quella di appartenenza.

4.2.4 Probabilità di ancoraggio

Una volta ottenuto l'output del classificatore è stato creato un nuovo dataset composto dai dati presenti nel test set aventi come classe di arrivo quella predetta dall'algoritmo. Questo dataset verrà utilizzato per determinare la probabilità di ancoraggio dell'esaminatore nella visualizzazione delle domande classificate come borderline che ammontano ad un totale di 35.
Per determinare una probabilità di ancoraggio, è stata scritta ed utilizzata una funzione che itera tra gli indici e le righe presenti nel dataset e, se la riga visionata in quel momento assume il valore di *border* nella colonna *Admission_state*, allora la funzione controlla e inserisce all'interno di una variabile quanti dei 10 valori subito precedenti in *Ground_truth* sono successivamente uguali tra loro: il numero di valori precedenti uguali ottenuti per la riga *border* presa in considerazione determina la probabilità di ancoraggio, che va da un minimo di 0 ad un massimo di 10. Le rispettive probabilità vengono poi inserite in una nuova colonna che prende il nome di *anchoring_prob.* Questo procedimento permette di comprendere quante domande di seguito sono state valutate positivamente (o anche negativamente) prima di quella borderline.
Al termine dell'operazione 85 istanze all'interno del dataset hanno una probabilità di ancoraggio uguale a 0, 19 istanze hanno una probabilità uguale a 1, 6 istanze hanno una probabilità uguale a 2, 4 istanze hanno una probabilità uguale a 3, un' istanza ha una probabilità uguale a 4, 2 istanze hanno una probabilità uguale a 5 e, infine, 3 istanze hanno una probabilità di ancoraggio uguale a 6.

La probabilità di ancoraggio permette di stabilire se ci si trova di fronte ad un possibile ancoraggio o meno, di conseguenza, una volta ottenute queste probabilità, è stata creata una nuova ed ultima colonna che prende il nome di *Anchoring_state* che assume valore 0 per le istanze con una probabilità di ancoraggio inferiore o uguale a 4, e valore 1 per le istanze con una probabilità di ancoraggio maggiore di 4 e che risultano essere, quindi, ancorate. Al termine del procedimento il dataset conta 115 istanze non ancorate e soltanto 5 istanze ancorate.

4.2.5 Random Forest Classifier

Individuato il possibile ancoraggio presente nei dati, è stato utilizzato un classificatore per osservarne la performance nel predire lo stato di ancoraggio delle istanze automatizzando il processo. In questo caso sono state escluse dal training set le seguenti features: *Anchoring_state* (essendo la classe di arrivo), *Admission_state* e *Ground_truth*. Le ultime due features sono state rimosse poiché non rilevanti nell'individuazione finale dello stato di ancoraggio, in particolare l'*Admission_state* dovrebbe essere dedotto dai valori presenti in *Admission_rate.*
Inoltre, anche in questo caso, è stato utilizzato il metodo SMOTE al fine di bilanciare le classi in modo da permettere all'algoritmo di apprendere le informazioni relative ad entrambe.

Il classificatore utilizzato prende il nome di Random Forest Classifier ed è un algoritmo che rientra tra gli *ensamble methods*: questo tipo di algoritmi utilizza solitamente una tecnica che prende il nome di *bootstrap*, che utilizza i dati presenti nel training set per creare diversi dataset generati in modo randomico della stessa dimensione di quello originale. Ognuno di questi dataset viene successivamente utilizzato per allenare un classificatore "base", aggregando infine le predizioni finali ottenute da ognuno. Gli ensamble methods hanno origine dall'idea che la conoscenza collettiva (quindi, in questo caso, quella di più classificatori assieme) superi quella del singolo.
Il Random Forest utilizza come classificatore base un albero di decisione (o Decision Tree): ogni bootstrap sample, creato attraverso il training set, viene utilizzato per allenare un albero di decisione differente, ogni albero, infatti, si allena su *n* attributi diversi selezionati in modo randomico. Infine, la predizione di questi alberi viene comparata e l'output finale viene stabilito attraverso un voto di maggioranza dei classificatori.
Anche in questo caso, prima di avviare l'algoritmo, sono stati ricercati i parametri migliori per i dati in input, trovando i seguenti parametri: *criterion="gini",max_depth=2, n_estimators=50,oob_score=True, random_state=0* .
Nella figura 10 è possibile vedere i risultati del classificatore attraverso l'utilizzo di una confusion matrix.

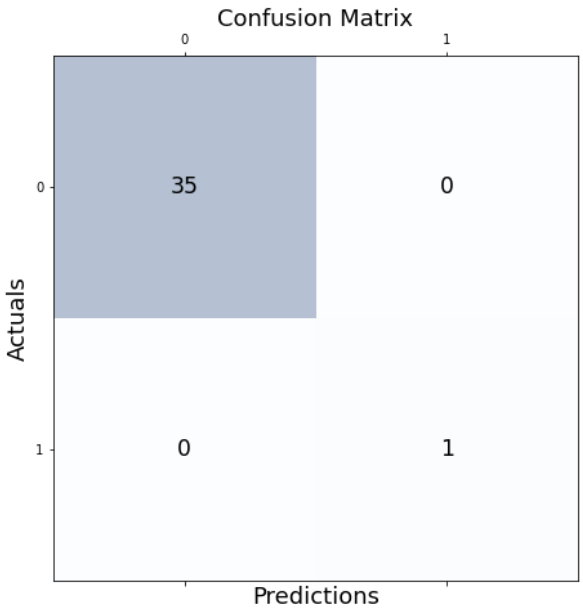

Figura 10: confusion matrix Random Forest Classifier.

Nella confusion matrix il valore 0 indica le istanze predette come non ancorate, mentre il valore 1 indica quelle ancorate. Attraverso la matrice di confusione risulta evidente che il classificatore abbia avuto una performance ottimale, con un accuracy pari a 1. In questo caso, quindi, tutti le istanze risultano classificate correttamente, con 35 punti classificati nella classe 0 ed un punto soltanto classificato nella classe 1.

4.2.6 Aprendo la scatola nera

Le operazioni effettuate precedentemente hanno permesso di rilevare quali delle istanze fossero, secondo il classificatore, ancorate. L'operazione successiva è quella di aggirare il problema etico osservato all'inizio del capitolo, che non permette l'applicazione di algoritmi a scatola nera in contesti delicati per la vita delle persone, poiché è impossibile spiegare del tutto come la decisione sia stata presa all'interno dell'algoritmo. In questo caso, al fine di spiegare la logica dietro il Random Forest Classifier e permetterne il suo utilizzo in un contesto simile, è stato utilizzato l'algoritmo di explainability LIME.

LIME (Local Interpretable Model-Agnostic Explanations) è un algoritmo model-agnostic, ovvero che non dipende dal modello che si vuole spiegare, in grado di estrapolare la logica della black box, utilizzandola come input. LIME spiega le

istanze aggiustando e cambiando i valori delle singole features e osservandone i relativi impatti sull'output.
In questo caso LIME è stato applicato sull'unica istanza classificata come *ancorata* dal Random Forest, osservandone così la spiegazione locale, visibile nella figura 11.

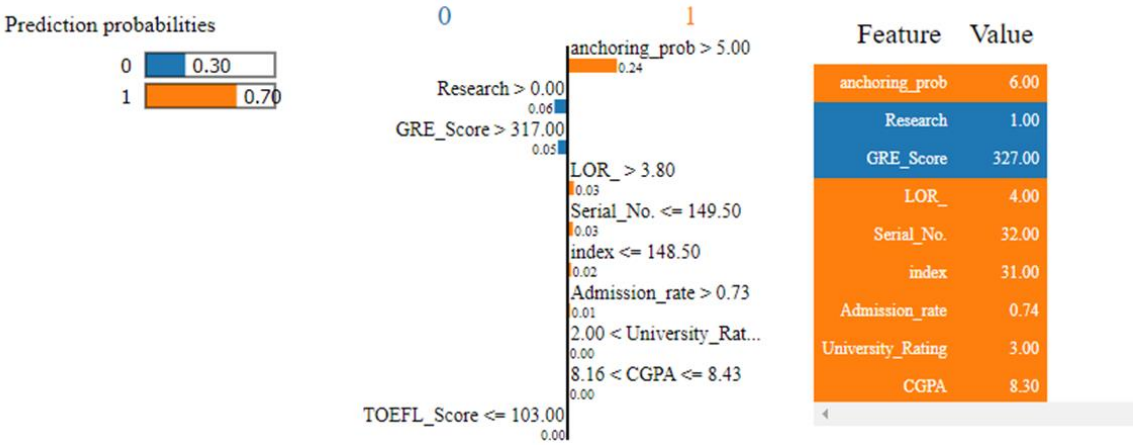

Figura 11: LIME output.

L'output generato da LIME mostra sulla sinistra la probabilità di appartenenza dell'istanza alla classe 1 (ancorata) o alla classe 0 (non ancorata): l'istanza viene classificata come ancorata con una probabilità del 70%, contro una probabilità di non ancoraggio del 30%. Sulla destra, invece, è possibile osservare come le features, e i loro rispettivi valori, hanno influenzato la predizione della classe: una probabilità di ancoraggio superiore a 5 appare, in questo caso, una delle caratteristiche più rilevanti per la classificazione dell'istanza.
L'utilizzo di LIME, su questo specifico dataset, ha l'obiettivo principale di mostrare il suo funzionamento ed applicazione, rendendo possibile così aggirare il problema etico legato al tema dell'intelligenza artificiale. Ovviamente, la probabilità di ancoraggio creata artificialmente risulta essere rilevante per l'algoritmo nella predizione della classe, lasciando intendere un corretto funzionamento dell'algoritmo. Purtroppo, in questo caso, il dataset utilizzato non è provvisto di marche temporali (timestamps) fondamentali per la rilevazione del bias di ancoraggio per il quale l'ordine temporale ha una grande importanza.

L'applicazione di LIME, o di qualsiasi altro algoritmo di explainability nella creazione di un modello permette, quindi, di evitare la presenza di bias al suo interno, rendendo comprensibile il suo funzionamento e consentendone così il suo utilizzo: se,

ad esempio, ci si trovasse davanti ad un modello con un bias di tipo razziale al suo interno, sarebbe possibile, attraverso LIME, vedere che la colonna riguardante l'etnia influenza l'algoritmo nella scelta della classe.

4.3 Secondo dataset: Movie Lens

Il secondo dataset[59] utilizzato prende il nome di MovieLens e contiene votazioni riguardanti diversi film da parte degli utenti. In questo caso il dataset verrà utilizzato per osservare un possibile ancoraggio prendendo in considerazione le votazioni effettuate da due diversi utenti: l'user ID numero 1 e il numero 6.

4.3.1 Prima osservazione dei dati

Il dataset è composto da 4 colonne e 100836 righe e, nella figura 12, è possibile osservarne le prime 5 righe.

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931

Figura 12: visualizzazione delle prime 5 righe del dataset MovieLens.

Come visibile nella figura sopra riportata, il dataset contiene le seguenti features:

-userId: numero che identifica l'utente.

-movieId: numero identificativo del film che l'utente ha votato.

-rating: punteggio da 1 a 5 assegnato al film a seconda dell'indie di gradimento dell'utente.

[59] https://grouplens.org/datasets/movielens/

-timestamp: marca temporale della votazione.

Anche in questo caso, il dataset non contiene valori mancanti in nessuna delle colonne.

4.3.2 Preparazione dei dati

Anche in questo caso, al fine di rilevare il possibile ancoraggio presente nei dati, è necessario aggiungere una colonna all'interno del dataset che sia in grado di fornirci un' "oggettiva" valutazione del film, in modo da poter osservare quali film votati dall'utente siano da considerare borderline.

Per ottenere una valutazione in qualche modo oggettiva dei film valutati dagli utenti, è stata creata ed inserita una nuova colonna che prende, anche in questo caso, il nome di *ground_truth* e contiene la media di tutte le votazioni effettuate dagli utenti per ogni specifico film. I valori sono stati successivamente arrotondati, per comodità e maggiore chiarezza, ad un massimo di due cifre.

Un ulteriore e fondamentale passaggio, possibile attraverso questo dataset, è quello di ordinare le valutazioni in base al loro ordine temporale utilizzando i timestamps: come già visto precedentemente, l'ordine in cui vengono presentate le istanze è fondamentale al fine di rilevare un possibile ancoraggio dal momento che l'utente viene influenzato, in questo caso, dalle votazioni date precedentemente alla valutazione di un film borderline.

Nella figura 13 è possibile osservare nuovamente le prime cinque righe del dataset ampliato ed ordinato:

	userId	movieId	rating	timestamp	ground_truth
66719	429	595	5.0	828124615	5.0
66716	429	588	5.0	828124615	5.0
66717	429	590	5.0	828124615	5.0
66718	429	592	5.0	828124615	5.0
66712	429	432	3.0	828124615	3.0

Figura 13: Prime cinque righe del dataset ampliato ed ordinato.

4.4 Primo utente

Il primo utente ad essere preso in considerazione per il rilevamento di un possibile ancoraggio è quello con *userId* uguale a 1. Per questo, è stato creato un nuovo dataset, a partire da quello già ampliato ed ordinato, contenente soltanto le valutazioni effettuate dall'utente 1: il nuovo dataset contiene 5 colonne e 232 righe.
Per una visione più chiara dei valori contenuti all'interno di questo dataset sono state osservate le valutazioni effettuate dall'utente, che risulta aver votato 124 film con il valore di 5.0, 76 film con il valore di 4.0, 26 film con il valore di 3.0, 5 film con il valore di 2.0 e soltanto un film con il valore di 1.0. L'utente ha quindi votato in modo maggiormente positivo i film proposti.
Inoltre, al fine di rilevare la probabilità di ancoraggio, è stata aggiunta un' ulteriore colonna che, in base al *rating* dato dall'utente, è in grado di indicare repentinamente se le valutazioni siano da considerare come positive o negative: la colonna prende il nome di *rating_value* ed assume il valore *high* se il rating fornito dall'utente per un determinato film ha un valore maggiore o uguale a 3.5, altrimenti assume il valore *low*.

	userId	movieId	rating	timestamp	ground_truth	rating_value
0	1	804	4.0	964980499	3.19	high
1	1	1210	5.0	964980499	3.46	high
2	1	2628	4.0	964980523	3.83	high
3	1	2826	4.0	964980523	3.50	high
4	1	2018	5.0	964980523	5.00	high

Figura 14: prime cinque righe del dataset userId 1 post-modifiche.

In questo caso la colonna *rating_value* assume lo stesso scopo della colonna *Ground_truth* creata nel primo dataset: permette, infatti, di osservare se le valutazioni effettuate precedentemente alla valutazione dell'istanza borderline, siano sequenzialmente maggiormente positive o negative, creando un possibile ancoraggio nell'utente. Contrariamente al primo dataset preso in considerazione, infatti, questo dataset fornisce sia dati che permettono di identificare chi ha effettuato la valutazione, sia dati che consentono di estrapolare una verità base attraverso la media delle diverse valutazioni di uno stesso film.

4.4.1 Probabilità di ancoraggio del primo utente

Per determinare la probabilità di ancoraggio è stata utilizzata una funzione che itera tra gli indici e le righe presenti nel dataset e, se la riga visionata in quel momento assume un valore maggiore o uguale a 3 e minore o uguale a 4 nella colonna *ground_truth*, allora il film viene considerato borderline e la funzione controlla e inserisce all'interno di una variabile quanti dei 10 valori subito precedenti in *rating_value* sono successivamente uguali tra loro: il numero di valori uguali, precedenti all'istanza borderline presa in considerazione nella prima parte della funzione, determina la probabilità di ancoraggio, che va da un minimo di 0 ad un massimo di 10. Le rispettive probabilità vengono poi inserite in una nuova colonna che prende il nome, anche in questo caso, di *anchoring_prob*. Questo procedimento permette di comprendere quante domande di seguito sono state valutate positivamente (o anche negativamente) dall'utente prima di quella in media considerata borderline.
Al termine dell'operazione 95 istanze all'interno del dataset hanno una probabilità di ancoraggio uguale a 0, 32 istanze hanno una probabilità uguale a 1, 18 istanze hanno una probabilità uguale a 2, 14 istanze hanno una probabilità uguale a 3, 7 istanze hanno una probabilità uguale a 4, 9 istanze hanno una probabilità uguale a 5, 3 istanze hanno una probabilità uguale a 6, un'istanza ha una probabilità uguale a 7, 6 istanze hanno una probabilità uguale a 8, 4 istanze hanno una probabilità uguale a 9 e, infine, 43 istanze hanno una probabilità di ancoraggio uguale a 10.

Ottenute queste probabilità, anche in questo caso, è stata creata una nuova ed ultima colonna che prende il nome di *Anchoring_state* che assume valore 0 per le istanze con una probabilità di ancoraggio inferiore o uguale a 4, e valore 1 per le istanze con una probabilità di ancoraggio maggiore di 4 e che risultano essere, quindi, ancorate. Al termine del procedimento il dataset conta 166 istanze non ancorate e 66 istanze ancorate.

4.4.2 Predizione del Random Forest Classifier

Individuate le istanze ancorate all'interno del dataset, il passaggio successivo è stato quello di utilizzare un classificatore per osservarne la performance nel predire lo stato di ancoraggio delle istanze automatizzando il processo. In questo caso sono

state escluse dal training set le seguenti features: *Anchoring_state* (essendo la classe di arrivo), *rating_value*, *ground_truth* e *anchoring_prob*. Le prime features sono state rimosse poiché non rilevanti nell'individuazione finale dello stato di ancoraggio, mentre la probabilità di ancoraggio è stata esclusa poiché in questo caso sono presenti altre feature di origine (e quindi non costruite) utili a dedurre l'ancoraggio, in particolare ad influire dovrebbe essere l'ordine in cui sono state valutate le istanze (ovvero il timestamp).
Inoltre, anche in questo caso, è stato utilizzato il metodo SMOTE al fine di bilanciare le classi in modo da permettere all'algoritmo di apprendere le informazioni relative ad entrambe: alla fine del bilanciamento sia la classe 0 che la classe 1 contengono 114 istanze.

Il classificatore utilizzato anche in questo caso è il Random Forest Classifier e i parametri migliori per i dati in input, risultano essere i seguenti: *criterion="gini",max_depth=8,n_estimators=300,oob_score=True,random_state=0*. Nella figura 15 è possibile vedere i risultati del classificatore attraverso l'utilizzo di una confusion matrix.

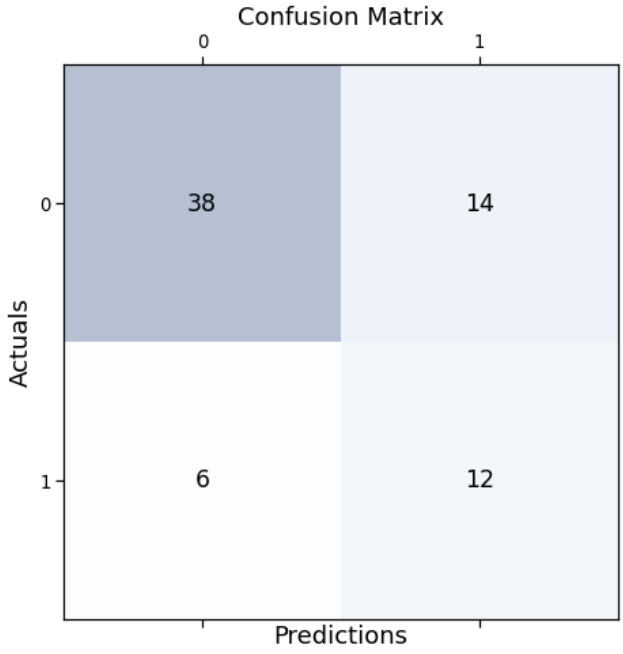

Figura 15: confusion matrix Random Forest Classifier.

Nella confusion matrix il valore 0 indica le istanze predette come non ancorate, mentre il valore 1 indica quelle ancorate. Attraverso la matrice di confusione risulta evidente

che il classificatore non abbia avuto una performance ottimale, con un accuracy pari a 0.71 ed una precisione uguale a 0.86 nel predire la classe 0 e pari soltanto a 0.46 nel predire la classe 1. Tuttavia, l'obiettivo non è quello di avere una performance ottimale da parte dei classificatori e, in questo caso, è anche normale che il classificatore faccia confusione non avendo in partenza i dati necessari per predire la classe.

4.4.3 Aprendo la scatola nera

Anche in questo caso, al fine di spiegare la logica dietro il Random Forest Classifier e permetterne il suo utilizzo in un contesto simile, è stato utilizzato l'algoritmo di explainability LIME.
LIME è stato applicato su una delle istanze classificate come *ancorate* dal Random Forest, ovvero quella di indice 68, osservandone così la spiegazione locale, visibile nella figura 16.

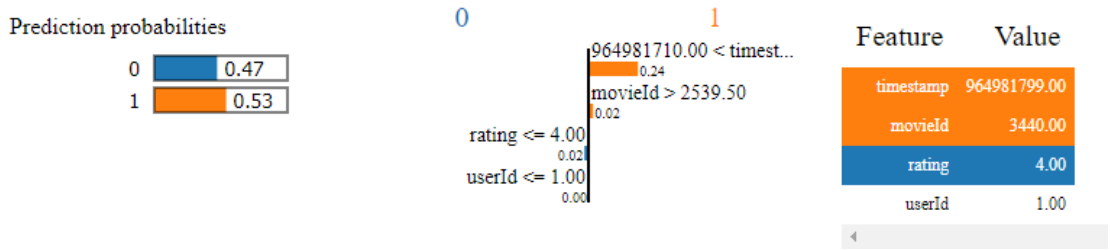

Figura 16: LIME output.

LIME permette di vedere che l'istanza viene classificata come ancorata con una probabilità del 53%, contro una probabilità di non ancoraggio del 47%. Sulla destra, invece, è possibile osservare come tra le features che hanno influenzato maggiormente la predizione della classe ci sia il timestamp, ovvero l'ordine in cui le istanze appaiono all'utente oltre all'identificativo del film. L'algoritmo sembra, dunque, aver afferrato in qualche modo la presenza di una connessione tra l'ordine in cui vengono presentate le istanze e l'ancoraggio.

4.5 Secondo utente

Il secondo ed ultimo utente ad essere preso in considerazione per il rilevamento di un possibile ancoraggio è quello con *userId* uguale a 6. La scelta di ripetere la

performance con due diversi utenti è utile al fine di vedere se questa metodologia può essere applicata anche su dati leggermente differenti, generalmente è sempre meglio ripetere più volte un esperimento al fine di testarne la sua efficacia. Dal momento che i procedimenti effettuati sono gli stessi utilizzati per il precedente utente, per l'ancoraggio del secondo utente ci si soffermerà meno sulle spiegazioni già fornite in precedenza. Anche in questo caso, è stato creato un nuovo dataset, a partire da quello già ampliato ed ordinato, contenente soltanto le valutazioni effettuate dall'utente 6: il nuovo dataset contiene 5 colonne e 314 righe.
L'utente risulta aver votato 40 film con il valore di 5.0, 102 film con il valore di 4.0, 152 film con il valore di 3.0, 13 film con il valore di 2.0 e 7 film con il valore di 1.0.
Anche qui è stata aggiunta la colonna *rating_value* che assume il valore *high* se il rating fornito dall'utente per un determinato film ha un valore maggiore o uguale a 3.5, altrimenti assume il valore *low*.

	userId	movieId	rating	timestamp	ground_truth	rating_value
0	6	592	3.0	845553109	3.50	low
1	6	590	5.0	845553109	5.00	high
2	6	380	4.0	845553110	4.15	high
3	6	296	2.0	845553110	2.00	low
4	6	150	4.0	845553110	3.00	high

Figura 17: prime cinque righe del dataset userId 6 post-modifiche.

4.5.1 Probabilità di ancoraggio del secondo utente

La probabilità di ancoraggio, nel dataset riguardante l'utente numero 6, è stata individuata allo stesso modo dell'utente numero 1: utilizzando una funzione in grado di calcolare e riempire con le probabilità di ciascuna istanza la nuova colonna *anchoring_prob*.
Al termine dell'operazione 90 istanze all'interno del dataset hanno una probabilità di ancoraggio uguale a 0, 97 istanze hanno una probabilità uguale a 1, 57 istanze hanno una probabilità uguale a 2, 27 istanze hanno una probabilità uguale a 3, 17 istanze hanno una probabilità uguale a 4, 12 istanze hanno una probabilità uguale a 5, 6 istanze hanno una probabilità uguale a 6, 6 istanze hanno una probabilità uguale a 7, un'istanza

ha una probabilità uguale a 8 e, infine, una sola istanza ha una probabilità di ancoraggio uguale a 10.

Ottenute le probabilità, anche in questo caso, è stata aggiunta la colonna *Anchoring_state* che assume valore 0 per le istanze con una probabilità di ancoraggio inferiore o uguale a 4, e valore 1 per le istanze con una probabilità di ancoraggio maggiore di 4 e che risultano essere, quindi, ancorate. Al termine del procedimento il dataset conta 288 istanze non ancorate e 26 istanze ancorate.

4.5.2 Predizione del Random Forest Classifier

Ottenuti i dati necessari per allenare il classificatore ed osservarne la performance nel predire lo stato di ancoraggio delle istanze, sono state escluse dal training set le seguenti features: *Anchoring_state* (essendo la classe di arrivo), *rating_value*, *ground_truth* e *anchoring_prob*.
Inoltre, anche in questo caso, è stato utilizzato il metodo SMOTE al fine di bilanciare le classi: alla fine del bilanciamento sia la classe 0 che la classe 1 contengono 202 istanze.

I parametri migliori del Random Forest Classifier per i dati in input risultano essere i seguenti: *criterion="gini",max_depth=8, n_estimators=300,oob_score=True, random_state=0.* Nella figura 18 è possibile vedere i risultati del classificatore attraverso l'utilizzo della confusion matrix.

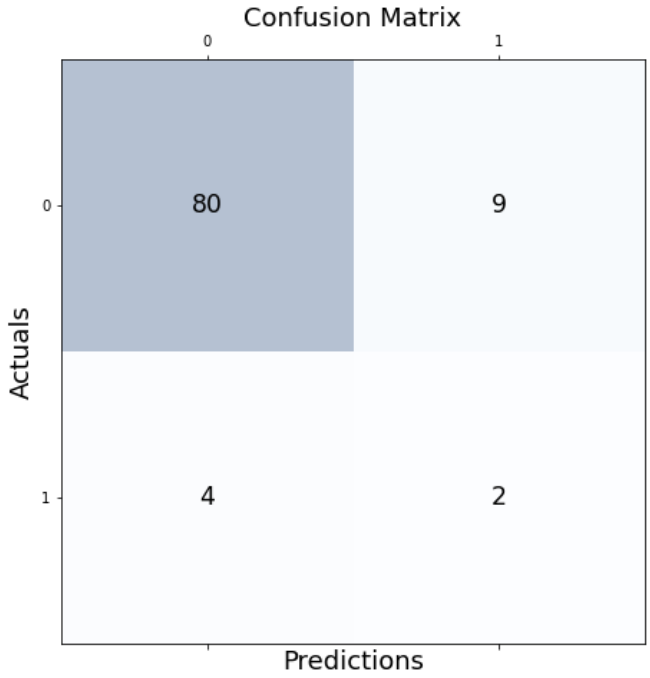

Figura 18: confusion matrix Random Forest Classifier.

Nella confusion matrix il valore 0 indica le istanze predette come non ancorate, mentre il valore 1 indica quelle ancorate. Anche in questo caso il classificatore non ha avuto una performance ottimale, con un accuracy pari a 0.86 e difficoltà nel predire la classe 1.

4.5.3 Aprendo la scatola nera

Questa volta, dal momento che la maggior parte dei punti è stata predetta correttamente nella classe 0, l'algoritmo di explainability LIME è stato utilizzato per osservare sia la logica dietro un'istanza classificata come ancorata, sia quella dietro un'istanza classificata come non ancorata.

Le istanze prese in considerazione sono quelle di indice 60 (classificata come non ancorata) e quella di indice 186, nelle figure 19 e 20 è possibile osservarne la spiegazione locale ottenuta attraverso LIME.

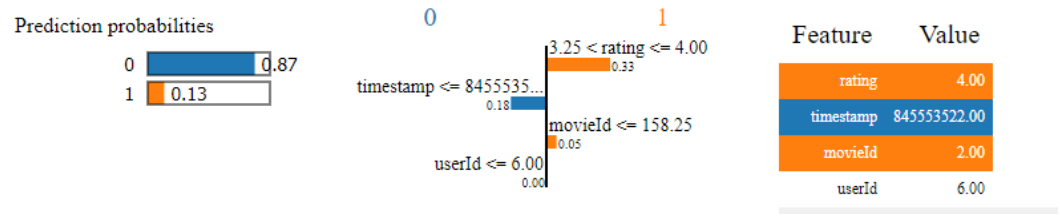

Figura 19: LIME output istanza 60.

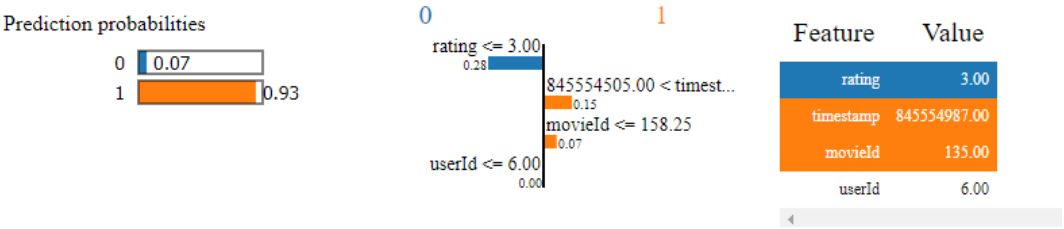

Figura 20: LIME output istanza 186.

Risulta evidente che l'algoritmo ritiene l'ordine in cui vengono valutate le istanze di fondamentale importanza per la predizione dell'ancoraggio: in figura 19, l'istanza con indice 60 viene classificata come non ancorata, con una probabilità di 0.87 di appartenenza alla classe 0 contro una probabilità di solo 0.13 di appartenenza alla classe 1. Al contrario, in figura 20, l'istanza con indice 186 viene classificata come ancorata, con una probabilità di appartenenza alla classe 1 pari a 0.93 ed una probabilità di appartenenza alla classe 0 pari soltanto a 0.07.

Il rilevamento di un possibile ancoraggio all'interno dei dati può essere utile al fine di ottenere una valutazione più oggettiva, ad esempio, ripassando al vaglio di una commissione le domande con un'alta probabilità di ancoraggio. Invece, nel caso delle valutazioni di prodotti o film si potrebbe, ad esempio, dare un peso minore sulla media finale alle recensioni che risultano essere ancorate.

4.6 Obiettivi e limitazioni

Il lavoro svolto, come già precedentemente sottolineato, si propone dunque di fornire un possibile approccio per l'individuazione del bias di ancoraggio, tentando

così di diagnosticarne la sua presenza all'interno dei dataset, piuttosto che concentrarsi su possibili metodi di mitigazione. L'obiettivo è, dunque, quello di esplorare e sviluppare metodologie che permettano la classificazione dell'ancoraggio attraverso algoritmi di intelligenza artificiale e, successivamente, osservare i metodi in grado di spiegarne il loro funzionamento.

Tuttavia, è importante tenere a mente anche le limitazioni presenti nello studio effettuato, come l'utilizzo di un dataset in gran parte costruito artificialmente al fine di raggiungere lo scopo proposto: determinate tipologie di dataset, infatti, sono più difficili da reperire online dal momento che contengono dati sensibili. Di conseguenza, il metodo utilizzato è legato ad ovvie limitazioni come la mancanza di una sequenza temporale nell'individuazione dell' ancoraggio.

CONCLUSIONE

L’utilizzo dell’algoritmo di explainability LIME ha dunque mostrato in modo efficiente la logica dietro le scelte prese dall’algoritmo del Random Forest Classifier, intuendo una correlazione tra l’ordine in cui le istanze vengono presentate e la classe di arrivo. Una metodologia simile può essere utilizzata per spiegare la scatola nera di algoritmi anche più complessi, consentendone il suo utilizzo in campi delicati come quello della medicina, e permettendo l’individuazione di bias cognitivi al suo interno, prima ancora del suo utilizzo.

Come già accennato precedentemente, l’individuazione del bias di ancoraggio all’interno dei dati, apre successivamente le porte a possibili mitigazioni del bias, permettendo un maggiore controllo sul peso che una valutazione ancorata può avere all’interno di una votazione o cambiando, ad esempio, l’ordine in cui le domande di ammissione vengono presentate ad un esaminatore. Un’ulteriore possibile mitigazione consiste nella sostituzione del valore ancorato con una media ottenuta attraverso diverse valutazioni.

Gli algoritmi di explainability rappresentano, dunque, un passo fondamentale per un’evoluzione tecnologica più equa, in grado di evitare ripercussioni negative in diversi contesti sociali anche attraverso l’utilizzo di algoritmi a scatola nera, ancor prima della loro distribuzione ed applicazione.

Lo studio presentato, ovvero la classificazione dell’ancoraggio presente nelle istanze e la sua successiva spiegazione attraverso l’algoritmo LIME, potrebbe essere quindi utilizzato al fine di validare algoritmi di ordinamento di dati sensibili a questo tipo di bias e aventi un impatto rilevante nella vita delle persone, con lo scopo di individuarlo e, successivamente, mitigarlo in maniera efficace e *human-centered*, ovvero sorpassando l’ostacolo etico e aprendo la black box.

BIBLIOGRAFIA

- A.TURING, *Computing Machinery and Intelligence*, 1950.
- BECHERA, H. DAMASIO, *The somatic marker hypothesis*, 2005
- BLESCHE-RECHEK, *Evolution and the trolley problem: people save five over one unless the one is young, genetically related, or a romantic partner,* 2010.
- D. KAHNEMAN, A. TVERSKY, *Judgment under uncertainty: Heuristics and biases*, 1982
- D. KAHNEMAN, A. TVERSKY, *Subjective probability: A judgment of representativeness*, 1972
- DEXE, J. FRANKE, U. SODERLUND, K. van BERKEL, N. JENSEN, R H. et al., *Explaining automated decision-making: a multinational study of the GDPR right to meaningful information*, 2022.
- E. J. LANGER, *The illusion of control*, 1975.
- FINKELSTEIN, L. BURKE, M. RAJU, *Age discrimination in simulated employment contexts: An integrative analysis*, 1995.
- H. BLACK, R. SCHWEITZER, M. LEWIS, *Discrimination in Mortgage Lending*, 1978.
- H. WAINER, H. L. ZWERLING, *Evidence That Smaller Schools Do Not Improve Student Achievement*, 2006
- J. BUOLAMWINI, T. GEBRU, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 2018.
- M. AMERI, L. SCHUR, D. KRUSE, *The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior*, 2017.

- M. G. HASELTON, D. NETTLE, *The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases*, 2006.
- M. GARCIA, *Racist in the machine*, 2017.
- M. GLADWELL, *Blink: The power of thinking without thinking*, 2006.
- S. ZACCARO, C. KEMP, , *Leader traits and attributes*, 2004.
- Y. MAJIMA, *Belief in pseudoscience, cognitive style and science literacy*, 2015.

SITOGRAFIA

- *AI and Empathy*: *Combining artificial intelligence with human ethics for better engagement*, luglio 2019, Pega, https://www.pega.com/system/files/resources/2019-11/pega-ai-empathy-study.pdf

- ALIPERTO, *Prestiti e machine learning, sul cammino i nodi privacy e black box,* 2022, CORCOM, https://www.corrierecomunicazioni.it/finance/prestiti-digitali-alla-prova-machine-learning-ma-vanno-sciolti-i-nodi-privacy-e-black-box/

- GHOLIPOUR, *We Need to Open the AI Black Box Before It's Too Late,* 2018, Futurism, https://futurism.com/ai-bias-black-box

- https://grouplens.org/datasets/movielens/

- https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

- *I 6 settori rivoluzionati da Intelligenza Artificiale e Machine Learning*, 2017, VIDIEMME, https://www.vidiemme.it/intelligenza-artificiale-e-machine-learning/

- *Intelligenza artificiale forte,*10 marzo 2021 ore 23:51, Wikipedia, https://it.wikipedia.org/wiki/Intelligenza_artificiale_forte

- J. DUSTIN, *Amazon scraps secret AI recruiting tool that showed bias against women*, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

- J. ECHTERHOFF, M.YARMAND, J. MCAULEY, *AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks*, 2022, https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

- J. LAURET, *Amazon's sexist AI recruiting tool: how did it go so wrong?*, 2019, https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

- N. KAYSER-BRIL, *Google apologizes after its Vision AI produced racist results*, 2020, ALGORITHM WATCH, https://algorithmwatch.org/en/google-vision-racism/

- O. SCHWARTZ, *Untold History of AI: Algorithmic Bias Was Born in the 1980s: A medical school thought a computer program would make the admissions process fairer-but it did just the opposite*, 2019, IEEE Spectrum, https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias#toggle-gdpr

- S. VALESINI, *Google Photos scambia afroameticani per gorilla*, 2015, WIRED, https://www.wired.it/attualita/tech/2015/07/02/google-photo-scambia-afroamericani-per-gorilla/

- SANKIN, D. MEHROTRA, S. MATTU, D. CAMERON, A. GILBERTSON, D. LEMPRES, J. LASH, *Crime Prediction Software Promised to Be Free of Biases*, dicembre 2021, GIZMODO, https://gizmodo.com/crime-prediction-software-promised-to-be-free-of-biases-1848138977

- *Storia dell'Intelligenza Artificiale, da Touring ai giorni nostri*, maggio 2019, Osservatori.net digital innovation, https://blog.osservatori.net/it_it/storia-intelligenza-artificiale

- T. DAVENPORT, *The potential for artificial intelligence in healthcare*, giugno 2019, PMC, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/

- T. SIMONITE, *When It Comes to Gorillas, Google Photos Remains Blind*, 2018, WIRED, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

- World Bank Group, *The World Bank in Gneder*, 11 ottobre 2022, https://www.worldbank.org/en/topic/gender/overview