POLITECNICO
MILANO 1863

Progetto di Ingegneria Informatica

Docente: Prof.ssa Mariagrazia Fugini

Responsabili: Prof.ssa Letizia Tanca, Prof. Fabio Schreiber, Dott.ssa Chiara Criscuolo, Dott. Tommaso Dolci

Alessandro Aldo Marina
Codice persona: 10708620
Matricola: 956172

Introduzione - l’etica e il suo impatto

Il concetto di computer ethics viene introdotto per la prima volta da Norbert Wiener, un professore del MIT, che nel 1950 con la pubblicazione del suo libro *The Human Use of Human Beings,* è diventato il fondatore dell’etica informatica ponendo le basi di questa disciplina che utilizziamo ancora oggi. In particolare Wiener nella sua pubblicazione ha evidenziato come l’integrazione della tecnologia informatica nella società avrebbe generato una rielaborazione della società stessa e delle sue dinamiche, definendo questo cambiamento come l’inizio di una nuova rivoluzione industriale. In questo libro sono stati identificati i compiti e le sfide che sarebbero stati inclusi in questa rivoluzione, come ad esempio l'adattamento dei lavoratori a cambiamenti sul posto di lavoro, l'implementazione di nuove leggi e regolamenti da parte dei governi e la necessità per i sociologi di studiare e comprendere i nuovi fenomeni sociali e psicologici. Nonostante oggi venga riconosciuto a Wiener l’impatto del suo operato, al tempo la disciplina della computer ethics era comunque rimasta inesplorata fino alla metà degli anni ’60, quando l’interesse per questo ambito cominciò ad aumentare.

Sono stati i primi crimini informatici ad attirare l’attenzione sui reali impatti che la tecnologia poteva avere. Durante quel periodo, Donn Parker, un informatico e ricercatore membro della ACM (Association for Computing Machinery), iniziò a catalogare i crimini informatici e pubblicò numerosi libri e articoli sull'etica dei computer. Tuttavia, poiché non esistevano leggi in quel momento per affrontare tali crimini e fermare tali attività non etiche, fu necessario un codice etico per i computer. Parker propose all'ACM di adottare un codice di etica per i suoi membri, e nel 1973, l'ACM lo nominò a capo di un comitato per creare tale codice. Il codice stabiliva che i membri dell'ACM dovessero impegnarsi a fare del loro meglio per evitare danni al pubblico, al governo, ai datori di lavoro e ai clienti, a mantenere la riservatezza delle informazioni personali, a rispettare i diritti di proprietà intellettuale e a utilizzare i computer in modo responsabile e professionale. Negli anni successivi l’interesse nei confronti di questa disciplina è continuato a crescere, ma nonostante questo, la rapidità con cui la tecnologia si stava evolvendo non permetteva alla società di far fronte a queste rivoluzioni. È proprio su questo distacco che si era creato tra società e progresso tecnologico che James Moor (docente e ricercatore di Filosofia Intellettuale e Morale al Dartmouth College) ha basato le sue ricerche. Nel 1985, Moor ha scritto un articolo intitolato "What is computer ethics" per l'edizione della rivista "Computers and Ethics". Edizione che diventò il numero di maggior vendita nella storia della rivista. Moor ha sostenuto che i computer sono macchine logiche in grado di eseguire operazioni altrimenti impossibili per gli esseri umani. Questo crea situazioni in cui non esistono norme morali adeguate per affrontarle. Moor ha definito queste situazioni "vuoti politici" o "vuoti concettuali”.

Un vuoto politico avviene quando non c’è un’autorità legale o sociale che regola l’uso dei computer in una determinata situazione, ad esempio chi è il responsabile dei danni causati da un errore di programmazione?

Un vuoto concettuale si verifica invece quando non si ha una chiara comprensione dei concetti morali applicabili all’uso dei computer, per esempio cosa significa essere giusti, onesti o rispettosi quando si parla di computer?

Moor propone così la definizione di etica informatica: L'analisi della natura e dell'impatto sociale dei computer e delle regole morali che dovrebbero guidare il loro uso. Inoltre afferma che questa disciplina ha tre obiettivi principali:

- Identificare e chiarire i vuoti politici e concettuali creati dai computer;
- Formulare e giustificare le politiche e i principi morali che dovrebbero governare l'uso dei computer;
- Promuovere l'educazione e la consapevolezza etica tra gli utenti e i professionisti del settore informatico.

Moor conclude che l'etica informatica è una disciplina importante e necessaria per affrontare le sfide poste dai rapidi progressi tecnologici. Egli invita i filosofi, i legislatori, i tecnici e il pubblico a collaborare per sviluppare un quadro etico adeguato e flessibile per il futuro dell'informatica.

L'evoluzione del concetto di etica è poi arrivata con Deborah Johnson, una filosofa statunitense che ha introdotto il concetto di progettare la tecnologia in modo etico fin dalle prime fasi di sviluppo, considerando l'impatto etico delle decisioni di progettazione e delle scelte tecnologiche sulla società e l'ambiente. Ha introdotto il concetto di "co-shaping" secondo cui la tecnologia non è solo un prodotto o una conseguenza della società, ma piuttosto la tecnologia e la società si plasmano reciprocamente. Si arriva quindi all'epoca attuale in cui il filosofo americano Richard De George nel 2003 sostiene che l'adozione della tecnologia non dovrebbe essere fatta in automatico, ma piuttosto deve essere prima giustificata dal fatto che questa tecnologia aiuti a servire le persone e la società, introducendo così un richiamo alla responsabilità sociale delle aziende e dei professionisti del settore tecnologico nell'utilizzare le loro risorse per creare un impatto positivo sulla società.

In sintesi, l'etica nell'intelligenza artificiale (IA) è diventata sempre più importante negli ultimi anni, poiché l'IA ha iniziato ad avere un impatto significativo sulla vita delle persone. Infatti la sua presenza in processi decisionali critici alla base della società come la valutazione del credito, le decisioni giuridiche e gli algoritmi di raccomandazione, ha fatto si che i ricercatori, e tutta la comunità in generale, concentrassero maggiormente la propria attenzione per far si che questi sitemi comprendessero il più possibile i valori etici, in modo da garantire che l'IA potesse essere utilizzata in modo equo, responsabile e sicuro mettendo al primo posto i diritti delle persone. [1] [2] [3]

Il cluster etico

Il cluster etico rappresenta l'insieme degli elementi etici che un sistema di intelligenza artificiale deve rispettare. Oltre a fornire un framework per valutare sotto il punto di vista etico l'IA, è utile anche come guida durante la fase di progettazione e sviluppo di questi sistemi.

In generale non rispettare i principi del cluster etico può aggravare i bias sociali già esistenti, se non addirittura introdurne di nuovi.

Un bias è un errore sistematico di giudizio o di interpretazione che può portare a un errore di valutazione o alla formulazione di una decisione poco oggettiva. Un esempio di bias negli algoritmi è quello di Amazon, la quale ha introdotto un algoritmo per la scansione automatica dei curricula per le assunzioni, ma ha poi scoperto che l'algoritmo considerava tutti i candidati femminili come non idonei. Questo rappresenta un esempio significativo di bias negli algoritmi, che può avere conseguenze negative per l'uguaglianza di genere. Il problema sorge dal fatto che l'algoritmo era stato addestrato su un numero molto maggiore di curricula maschili rispetto a quelli femminili, il che ha portato ad una mancanza di rappresentatività del campione di dati. Inoltre, l'algoritmo preferiva volutamente i candidati provenienti da università prestigiose, dove la presenza maschile era ancora una volta maggiore rispetto a quella femminile. Questo ha contribuito ad aumentare il bias di genere nell'algoritmo e ad amplificare gli effetti negativi sulle donne che cercavano lavoro. È quindi chiaro come la prevenzione e l'individuazione di bias debba avere una priorità assoluta nella progettazione di questi sistemi

Il cluster etico comprende diversi principi etici tra cui: fairness, trasparenza, diversità, trust e privacy. In questo articolo ci concentreremo sull'analisi di uno di questi elementi: la fairness. [4] [5]

Fairness

Si riferisce alla necessità di progettare, sviluppare e utilizzare l'IA in modo da garantire un trattamento giusto per tutte le persone, senza alcuna forma di pregiudizio o discriminazione.

Esistono diverse tipologie di fairness che si possono applicare in base alle necessità progettuali, una parte importante nella progettazione di un algoritmo fair è dunque rappresentata dalla scelta della tipologia di fairness da utilizzare, infatti è difficile implementare allo stesso tempo più di una definizione di fairness poiché c'è un trade-off tra gli aspetti che ogni metrica punta a massimizzare. La prima categoria di fairness è quella di Group fairness con le sue relative metriche.

Per capire le metriche di Group fairness, è necessario avere una breve comprensione del funzionamento degli algoritmi che utilizzano queste metriche. Fondamentalmente, esistono diverse tipologie di algoritmi, l'algoritmo che abbiamo utilizzato in questo studio appartiene alla famiglia degli algoritmi predittivi che hanno il compito di prevedere una specifica caratteristica di ogni individuo basandosi sui valori noti di tutte le sue altre caratteristiche. Ad esempio, un algoritmo può essere utilizzato per prevedere se uno studente ha superato il test di ammissione all'università, basandosi sui suoi dati anagrafici e sulla media delle scuole superiori. L'obiettivo delle metriche di Group fairness è garantire che l'algoritmo fornisca previsioni giuste e imparziali per tutti gli individui, indipendentemente dalle loro caratteristiche personali. Nell'esempio precedente

vogliamo che l’algoritmo basi la sua previsione sulla media delle scuole superiori ma non sul sesso anagrafico.

Questa tipologia di fairness contiene al suo interno diverse definizioni che fanno riferimento al concetto di gruppo. La prima è la Statistical Parity che richiede che la percentuale di risultati positivi sia la stessa per tutti i gruppi di persone considerati. Ad esempio, se il 70% dei candidati di un gruppo considerato "svantaggiato" viene ammesso a una scuola, la metrica di Statistical Parity richiede che anche il 70% dei candidati del gruppo considerato "privilegiato" venga ammesso, se questi due gruppi hanno le stesse qualifiche (stesse caratteristiche tranne quelle che determinano l’appartenenza ad un gruppo o ad un altro). Un’altra definizione di Group Fairness nota come Predictive Parity impone che la probabilità che un individuo etichettato con un valore predetto positivo dall'algoritmo (ad esempio, il valore 1 indica che l'individuo ha passato il test), appartenga effettivamente al gruppo positivo (ad esempio, persone che hanno effettivamente passato il test), sia uguale tra i diversi gruppi.

L’altra tipologia di fairness è l’individual fairness la quale richiede che individui simili in base alle loro caratteristiche rilevanti ricevano lo stesso trattamento indipendentemente dal gruppo di appartenenza o da caratteristiche che non siano rilevanti per la decisione in questione. In questo modo si cerca di evitare la creazione di disparità individuali, non solo tra gruppi, ma anche all’interno di essi. [6]

Successivamente, sarà condotta un'analisi retrospettiva su ciascun dataset al fine di rilevare la presenza di eventuali bias. Dato che il nostro studio riguarda la verifica della fairness, e non la sua implementazione, utilizzeremo diverse metriche per ottenere risultati più precisi e considerare i diversi aspetti dei dati. Per entrambi i dataset presenti verrà prima descritto il dataset utilizzato e il preprocessing applicato, successivamente verrà mostrata la fase di processing dei dati, e infine verranno presentati e valutati i risultati ottenuti utilizzando delle metriche per decretare se è presente la fairness nel dataset in oggetto.

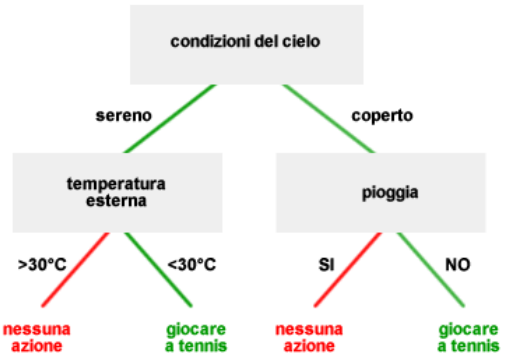
Strumenti di analisi

L'analisi del dataset è stata condotta interamente servendosi del Google Colab Notebook che è un ambiente di sviluppo integrato che consente di eseguire del codice in Python sul browser web e trova la sua applicazione nell'ambito della data science. In particolare ci siamo serviti della libreria sklearn di Python per la parte di processing. Tale libreria fornisce un insieme di algoritmi e strumenti avanzati per la trasformazione di dati tra cui l'algoritmo di machine Learning che abbiamo impiegato cioè il RandomForestClassifier.

Il Machine Learning è un sottoinsieme dell'IA che si occupa di creare sistemi che apprendono e migliorano le loro performance in base ai dati che utilizzano, e ha come obiettivo quello di imparare da esempi che gli vengono forniti (dati di training) per poi generalizzare e fare previsioni sui nuovi dati che gli vengono sottoposti (dati di test). I dati con cui opera l'algoritmo, indipendentemente dalla tipologia (dati di test o dati di training) hanno diversi attributi che rappresentano le caratteristiche che questi dati possiedono, come data di nascita, città di residenza o voto di laurea. Questi attributi si dividono per tipologia in base all'uso che ne fa l'algoritmo e le tipologie tipiche di questi attributi sono:

- Attributi di input: sono le variabili indipendenti o predittive che vengono utilizzate come punto di partenza per fare una previsione o una classificazione;
- Attributi di output: variabile dipendente o la variabile target che l'algoritmo cerca di prevedere o classificare;
- Attributi di peso: assegnano un peso diverso ad ogni attributo di input in base all'importanza che ogni attributo ha nella previsione del risultato.

Il RandomForestClassifier deve il suo nome al fatto che sfrutta diversi alberi di decisione che poi vengono fusi insieme per la classificazione dei dati, il cui l'obiettivo è predire la classe di appartenenza di un elemento in base alle sue caratteristiche. Un albero decisionale è una struttura che ha il compito di assegnare un'etichetta agli elementi che gli vengono dati in input, ed è composto da nodi, rami e foglie. Le foglie sono la parte terminale dell'albero e rappresentano le etichette finali assegnate agli elementi. I nodi sono punti in cui l'albero si ramifica in più percorsi (rami) che portano ad altri nodi o foglie, e ogni nodo rappresenta una scelta da fare basata su una caratteristica dei dati.



Esempio di albero decisionale

Il RandomForestClassifier viene preferito rispetto all'utilizzo di un solo albero decisionale in quanto il singolo albero ha la tendenza ad adattarsi eccessivamente ai dati di training perdendo la capacità di generalizzare (questo fenomeno è definito come overfitting di un modello). Dunque questo algoritmo, utilizzando diversi alberi decisionali, ciascuno allenato con subset diversi di dati, riesce a risolvere il problema di overfitting.[7]

Dataset: SAT Students

Il dataset [1][2] selezionato per questo esperimento contiene i dati di 1000 studenti, relativi alla media di ciascuno di essi e al punteggio che questi hanno ottenuto nella parte verbale e di matematica nel test SAT. Il test SAT è un test standardizzato in America riconosciuto per l'ammissione ai college nei vari stati. In particolare il dataset proviene da un'organizzazione chiamata Educational Testing Service, e presenta i seguenti attributi per ciascuno studente: sesso anagrafico (sex), percentile della parte verbale del SAT (sat_v), percentile della parte matematica del SAT (sat_m), somma dei percentili della parte verbale e matematica (sat_sum), media dei voti delle scuole superiori (hs_gpa), media dei voti del primo anno delle superiori (fy_gpa). I valori che questi attributi possono assumere sono riportati di seguito:

- sex: numero (1 o 2) in base al sesso
- sat_v: essendo un percentile assume valori da 0 a 100
- sat_m: segue la scala di sat_v
- hs_gpa: è la media dei voti su scala americana e varia da 0,0 a 4,0
- fy_gpa: segue la stessa scala di hs_gpa

Nella tabella sottostate sono riportate le prime 4 righe del dataset

	sex	sat_v	sat_m	sat_sum	hs_gpa	fy_gpa
0	1	65	62	127	3.40	3.18
1	2	58	64	122	4.00	3.33
2	2	56	60	116	3.75	3.25
3	1	42	53	95	3.75	2.42

[1] https://www.openintro.org/data/index.php?data=satgpa

[2] https://drive.google.com/file/d/1Y1ub6qj-EnhaLykWQFkQCLTUInEHIvCY/view?usp=sharing

Preprocessing del dataset

Nella fase di preprocessing stati mappati dei valori per rendere più semplice l'analisi. I valori dell'attributo "sex" sono stati mappati nei valori [0, 1], in modo da seguire lo standard di rappresentazione dei dati per questa tipologia. I valori relativi agli attributi sat_sum e hs_gpa sono stati mappati in 0 e 1, a seconda che il dato del singolo studente fosse maggiore o minore rispetto alla media generale dei valori dell'attributo corrispondente, quindi il valore 1 nell'attributo sat_sum indicherà che lo studente ha un voto maggiore della media mentre il valore 0 che lo studente ha un voto minore della media. Volendo porre l'attenzione unicamente sul fatto che uno studente sia sopra o sotto la media è stata adottata questa semplificazione con relativa perdita di informazione dei dati ma comunque compatibile con il nostro studio. Il mapping non è stato applicato all'attributo fy_gpa dato che non è una variabile target del nostro studio. Inoltre la scelta di lasciare inalterati i valori dell'attributo fy_gpa ha il vantaggio di avere un valore su una scala più precisa (0,0 - 4,0 anziché 0 - 1), attributo che l'algoritmo di machine learning Random Forest utilizza, insieme agli altri, per predire il valore dell'attributo sat_sum.

È stata inoltre condotta una pulizia e sanitization dei dati per rimuovere le tuple con valori nulli o non validi la quale non ha portato alla rimozione di tuple in quanto rispettavano tutte i requisiti.

Successivamente è stato applicato uno Standard Scaler il quale adotta una tecnica di normalizzazione dei dati che trasforma i dati in modo che abbiano valore atteso pari 0 e una deviazione standard unitaria. Questo fa si che l'algoritmo lavori con numeri distribuiti sulla stessa scala così da aumentare le prestazioni dello stesso, in questo caso le scale erano già simili ma non identiche e per questo è stata applicata questa trasformazione.

Come ultima fase di preprocessing i dati sono stati suddivisi in due gruppi: privilegiati e discriminati, sulla base del sesso anagrafico, che si presume essere il fattore su cui si basa la disparità. Questo è stato fatto per verificare se le metriche decisionali dell'algoritmo differiscono tra i due gruppi. Infine è stato individuato un attributo chiamato legittimo, in questo caso hs_gpa, il quale è considerato di rilievo per predire il valore dell'attributo target, in quanto è quello più significativamente correlato all'esito del SAT.

```
[ ]  #save the two groups indexes
     discriminated = []
     privileged = []
     for idx, i in enumerate(sensible_indexes):
       if i==sensible_values[0]:
         discriminated.append(sensible_indexes.index[idx])
       else:
         privileged.append(sensible_indexes.index[idx])
```

Divisione in gruppo privilegiato e discriminato

```
df['sat_sum'] = np.where(df['sat_sum'] < 104, 0, 1)
```

Codice per formattazione condizionale dei valori dell'attributo sat_sum. Se il valore nella tupla è minore della media (104) assegna 0, altrimenti 1.

Processing del dataset

Terminata la fase di preprocessing i dati vengono passati all'algoritmo il quale calcola il valore dell'attributo target basandosi sui dati di training. Il dataset che viene sottoposto all'algoritmo è diviso in due porzioni: il dataset di training e il dataset di test. I dati di train nel nostro caso costituiscono il 70 % (700 studenti) dei dati totali. I dati di addestramento (train) sono utilizzati per primi dall'algoritmo di RandomForestClassification e servono per apprendere la correlazione che c'è tra l'attributo target (sat_sum) e gli attributi input che invece rappresentano le caratteristiche dello studente.

Terminata la fase di apprendimento, l'algoritmo analizza il dataset di test e partendo dagli stessi attributi di input del training set ha il compito di prevedere i valori dell'attributo target.

Infine vengono comparati i valori dell'attributo target calcolati dall'algoritmo con quelli effettivamente presenti nel test set, e grazie a questo confronto è possibile stabilire l'accuratezza dei modelli predittivi dell'algoritmo stesso. Nella fattispecie l'accuratezza viene analizzata utilizzando un test diagnostico, nel quale sono presenti 4 gruppi dovuti al fatto che i valori della variabile target sono binari:

- Veri positivi (True Positives, TP): sono le previsioni positive corrette, cioè i casi all'interno del test set identificati come positivi dall'algoritmo che sono effettivamente positivi;
- Falsi positivi (False Positives, FP): sono le previsioni positive sbagliate, cioè i casi all'interno del test set identificati come positivi dall'algoritmo che sono in realtà negativi;
- Veri negativi (True Negatives, TN): sono le previsioni negative corrette, cioè i casi all'interno del test set identificati come negativi dall'algoritmo che sono effettivamente negativi;
- Falsi negativi (False Negatives, FN): sono le previsioni negative sbagliate, cioè i casi all'interno del test set identificati come negativi dall'algoritmo che sono in realtà positivi;

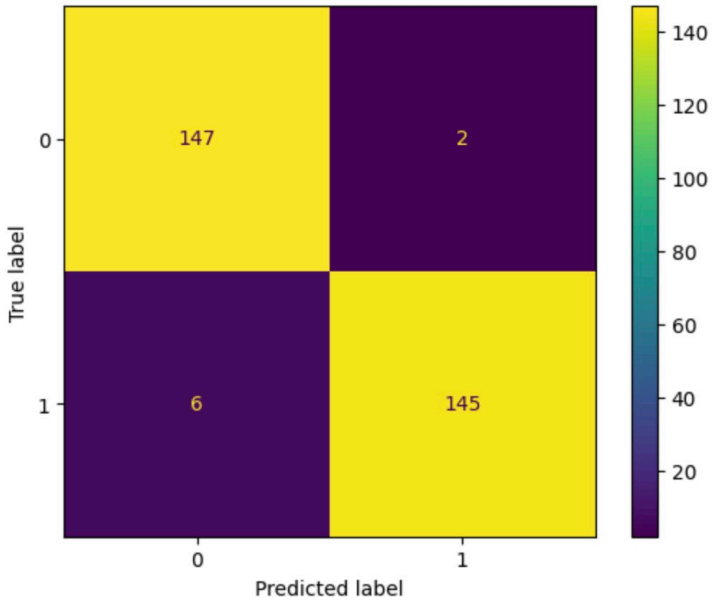
Un algoritmo accurato è in grado di classificare correttamente i dati forniti, fornendo risultati affidabili che possono essere utilizzati per prendere decisioni informate. Dunque l'accuratezza dell'algoritmo è l'elemento alla base della nostra analisi sui bias e la fairness del dataset.

```
#save the two groups indexes
Y = df[target_variable]
X = df.drop(target_variable, axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=1
```

Ripartizione in train set e test set

Esiti del processing

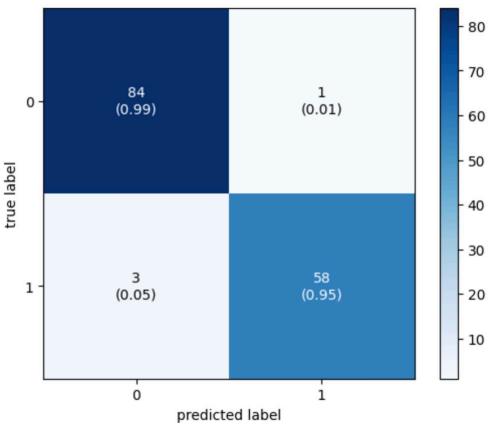
I dati sono mostrati utilizzando la confusion matrix che rappresenta i 4 gruppi del test diagnostico di cui sopra. Di seguito è riportata la matrice relativa al dataset:



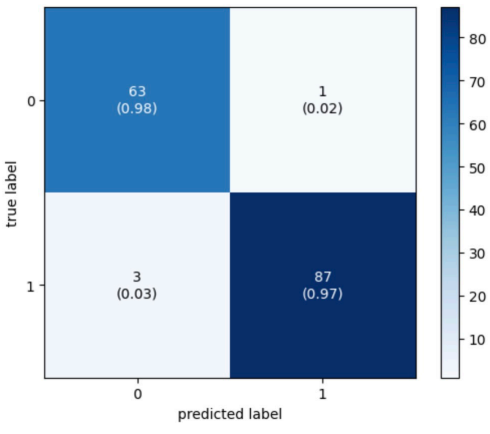
La zona della matrice che deve presentare il maggior numero di dati affinché l'accuratezza sia elevata è quella della diagonale principale, che rappresenta i TP e TN. In questo caso l'accuratezza dell'algoritmo è del 97% e la matrice riporta il 97% dei casi sulla diagonale principale. L'accuratezza è calcolata come segue:

- True Positive (145)
- True Negative(147)
- False Positive (6)
- False Negative(2)
- Accuratezza = (TP+TN)/(TP+TN+FP+FN) = 0.973 = 97.33%.

Tuttavia sapere che l'algoritmo è accurato non ci fornisce informazioni riguardo alla presenza di bias e alla fairness, infatti è necessario sottoporre i dati ottenuti ad un'ulteriore analisi. Vengono dunque calcolate due confusion matrix corrispondenti al gruppo privilegiato (costituito da uomini) e discriminato (costituito da donne) per condurre uno studio incrociato su di essi e per confrontare le performance dell'algoritmo nei due rispettivi casi. Le due confusion matrix sono le seguenti:



Discriminato



Privilegiato

Fairness Metrics Computation

L'ultima fase di questo studio del dataset si concentra sulla computazione e sulla verifica della fairness del dataset e per farlo vengono calcolati 10 valori per ciascuno dei due gruppi e vengono messi poi a rapporto per evidenziare con facilità le differenze tra i due. I parametri che vengono confrontati seguono la relativa notazione:

- d è il valore predetto della variabile target;
- Y è il valore effettivo della variabile target nel dataset;
- G è l'attributo protetto che indica a quale dei due gruppi appartiene il dato, privilegiato (priv) o discriminato (discr);
- L è l'attributo legittimo utilizzato per il calcolo della parità statistica condizionale.

La lista di metriche impiegate nel confronto invece è la seguente:

1. Group Fairness: (d=1|G=priv) = (d=1|G=discr)
2. Conditional Statistical Parity: (d=1|L=l, G=priv) = (d=1|L=l,G=discr)
3. Predictive Parity: (Y=1|d=1,G=priv) = (Y=1|d=1,G=discr)
4. Predictive Equality: (d=1|Y=0,G=priv) = (d=1|Y=0,G=discr)
5. Equal Opportunity: (d=0|Y=1,G=priv) = (d=0|Y=1,G=discr)
6. Equalized Odds: (d=1|Y=i,G=priv) = (d=1|Y=i,G=discr), i $\in$ 0,1
7. Conditional UseAccuracy Equality: (Y=1|d=1, G=priv) = (Y=1|d=1,G=discr) e (Y=0|d=0,G=priv) = (Y=0|d=0,G=discr)
8. Overall Accuracy Equality: (d=Y, G=priv) = (d=Y, G=priv)
9. Treatment Equality: (Y=1, d=0, G=priv)/(Y=0, d=1, G=priv) = (Y=1, d=0, G=discr)/(Y=0, d=1, G=discr)
10. FOR Parity: (Y=1|d=0, G=priv) = (Y=1|d=0,G=discr)

Di seguito vengono spiegate più nel dettaglio le diverse metriche. In particolare quando si parla di "ricevere un esito positivo (negativo)" ci si riferisce al fatto che l'algoritmo, nella sua previsione, abbia assegnato il valore 1 (0) all'attributo target. Quando invece ci si parla di "avere effettivamente una decisione positiva (negativa)" ci si riferisce al vero valore dell'attributo target presente nel dataset cioè 1 (0).

- Group Fairness: richiede che gli individui del gruppo discriminato e privilegiato abbiano la stessa probabilità di ottenere un esito positivo;
- Condizional statistical Parity: richiede che gli individui del gruppo discriminato e privilegiato, con gli stessi valori per gli attributi legittimi L, abbiano la stessa probabilità di avere un esito positivo;
- Predictive Parity: richiede che gli individui discriminati e privilegiati che ricevono un esito positivo abbiano la stessa probabilità di effettivamente avere una decisione negativa;
- Predictive Equality: richiede che individui discriminati e privilegiati con una decisione negativa abbiano la stessa probabilità di ricevere una predizione positiva. In questo caso uomini e donne che hanno un voto del SAT sotto la

media devono avere la stessa probabilità di ricevere un esito positivo come previsione dell'algoritmo;

- Equal Opportunity: richiede che gli individui discriminati e privilegiati che hanno effettivamente decisione positiva abbiano la stessa probabilità di ricevere una previsione negativa;
- Equalized Odds: è la combinazione di Predictive Equality e Equal Opportunity, richiede che gli individui discriminati e privilegiati che hanno una decisione effettivamente positiva abbiano la stessa probabilità di ricevere una previsione positiva. Allo stesso modo, i gruppi discriminati e privilegiati che hanno effettivamente decisione negativa dovrebbero avere la stessa probabilità di ricevere previsione positiva;
- Conditional Use Accuracy Equality: è la combinazione di Predictive Parity e di FOR Parity, richiede che gli individui discriminati e privilegiati che hanno una decisione effettivamente positiva abbiano la stessa probabilità di ricevere un risultato positivo come previsione. Allo stesso modo, gli individui discriminati e privilegiati con una decisione negativa dovrebbero avere la stessa probabilità di ricevere una previsione con risultato negativo;
- Overall Accuracy Equality: richiede che gli individui discriminati e privilegiati abbiano la stessa probabilità di ricevere una decisione positiva e una previsione positiva. Dovrebbero inoltre avere la stessa probabilità di ricevere una decisione negativa e una previsione negativa. In questo caso gli uomini e le donne devono avere la stessa probabilità che avendo media del SAT inferiore ricevano una previsione secondo cui hanno una media inferiore.
- Treatment Equality: richiede che il rapporto tra le previsioni errate e corrette sia lo stesso per individui discriminati e privilegiati;
- FOR Parity: richiede che gli individui discriminati e privilegiati che sono previsti per avere un risultato negativo abbiano la stessa probabilità di ottenere un esito positivo (nel dataset). In altre parole è la probabilità che un membro di un gruppo particolare non riceva un esito negativo quando avrebbe dovuto riceverlo in base alle sue caratteristiche. In questo caso deve esserci la stessa percentuale di donne e di uomini che hanno ricevuto la previsione di 1 nell'attributo target (SAT sopra la media) nonostante basandosi sulle loro caratteristiche l'algoritmo avrebbe dovuto assegnargli il valore 0 (SAT sotto la media).

Tutti questi parametri calcolati per ognuno dei due gruppi sono stati messi a rapporto tra loro portando al risultato nella tabella della pagina successiva.

È necessario scegliere un intervallo entro cui la differenza tra il valore della metrica del gruppo privilegiato e del gruppo discriminato è accettabile. In questo test è stato scelto come intervallo +/- 0.1. Questo significa che tutti i valori nella colonna "Value" che appartengono all'intervallo [0, 9] indicano che il gruppo discriminato soffre di unfairness, mentre quelli che appartengono al range [1.11, ∞] indicano che il gruppo privilegiato soffre di unfairness, e per quanto riguarda i valori che ricadono tra [0.9, 1.1] indicano che non è presente una differenza di trattamento tra i due gruppi. Esaminando i risultati di questa analisi osserviamo

	Metric	Value	Value_discriminated_group	Value_privileged_group
0	GroupFairness	0.707192	0.404110	0.571429
1	ConditionalStatisticalParity	0.692308	0.500000	0.722222
2	PredictiveParity	0.994350	0.983051	0.988636
3	PredictiveEquality	0.752941	0.011765	0.015625
4	EqualOpportunity	0.677778	0.049180	0.033333
5	EqualizedOdds	0.740598	0.983607	0.752941
6	ConditionalUseAccuracyEquality	1.005780	0.994350	1.011494
7	OverallAccuracyEquality	0.888889	0.666667	1.333333
8	TreatmentEquality	1.054795	438.000000	462.000000
9	FORParity	1.318182	0.034483	0.045455
10	FN	0.948052	0.020548	0.019481
11	FP	1.054795	0.006849	0.006494

che sono presenti diverse metriche che non rientrano nell'intervallo considerato accettabile, cioè tutte le metriche con valori inferiori a 0.9, o superiori a 1.1:

- Group Fairness
- Conditional Statistical Parity
- Predictive Equality
- Equal Opportunity
- Equalized Odds
- FOR Parity

Il valore di Conditional Statistical Parity si discosta di 0.31 dal valore ideale (1) e indica che gli individui del gruppo discriminato hanno una probabilità inferiore di ottenere un esito positivo rispetto a quelli del gruppo privilegiato a parità di valore dell'attributo legittimo. Questa metrica però tiene in considerazione il numero totale di Positivi per ogni gruppo divisi per il numero di studenti con un valore specifico di attributo legittimo. In questo caso nonostante il numero di Positivi sia molto simile (141 discriminato e 150 privilegiato) il numero di studenti con il valore 1 nell'attributo legittimo del gruppo discriminato è 101 mentre nel gruppo privilegiato è 40, ed è per questo che è presente una grande differenza nei valori dei due gruppi. La metrica di Group Fairness si discosta di 0.3 e indica che il gruppo discriminato ha una probabilità inferiore di ottenere un esito positivo rispetto al gruppo privilegiato . La metrica di Predictive Equality è anch'essa fuori dal range di valori accettabile e indica che gli studenti del gruppo discriminato con un valore effettivamente negativo (SAT sotto la media) hanno una probabilità inferiore di ricevere una predizione positiva rispetto agli studenti del gruppo privilegiato con un valore effettivamente negativo di "sat_sum". Equal Opportunity si discosta di 0.33 e indica che gli studenti del gruppo discriminato hanno una possibilità maggiore di essere predetti avere un voto del sat sotto la media nonostante abbiano in realtà avuto un voto sopra la media. Equalized Odds si discosta di 0.26 e indica che gli individui del gruppo discriminato che sono effettivamente positivi hanno una probabilità inferiore di ricevere una previsione positiva rispetto agli stessi individui positivi del gruppo privilegiato, così come gli individui discriminati che hanno un valore negativo hanno una minore

probabilità di avere una predizione negativa rispetto al gruppo privilegiato. Questo indica che l'accuratezza nell'individuare correttamente gli individui nel gruppo discriminato è inferiore rispetto al gruppo privilegiato. FOR Parity si discosta di 0.32 sopra il valore ideale. Analizzando i valori del FOR Parity per ciascun gruppo vediamo che il gruppo discriminato ha un valore di 0.034 mentre il gruppo privilegiato ha un valore di 0.045, questo valore è calcolato come FN / Negativi totali. Questo significa che il gruppo privilegiato ha percentualmente una probabilità maggiore di ricevere una previsione negativa nonostante secondo le sue caratteristiche dovrebbe averne una positiva.

Considerando queste metriche notiamo che il valore della metrica FOR Parity è l'unico a indicare una discriminazione nei confronti del gruppo privilegiato, questo è dovuto al fatto che il numero di TN è minore per questo gruppo rispetto a quello discriminato e quindi anche a parità di FN individuati (3) abbiamo una percentuale di FN individuati sul totale dei negativi maggiore. Dunque prendendo in considerazione questi aspetti, il valore di questa metrica non ci fa ipotizzare che ci sia unfariness nei confronti del gruppo privilegiato. Osservando invece le altre 5 metriche vediamo che ci forniscono dei valori a supporto dell'ipotesi che il gruppo etichettato inizialmente come discriminato lo sia realmente.

In conclusione essendoci complessivamente 5 metriche a supporto del fatto che nel dataset sia presente unfairness nei confronti del gruppo discriminato ed essendo la media dei valori di queste metriche di 0.29 sotto il valore ideale (1), siamo portati a sostenere che sia effettivamente presente unfairness nei confronti del gruppo discriminato, cioè quello in cui gli studenti sono di sesso femminile.

Esperimento 2: Portuguese students

Questo secondo dataset [3][4] proviene dalla UCI Machine Learning Repository e riguarda il rendimento di 649 studenti nelle scuole secondarie, provenienti da due istituti portoghesi. Gli attributi dei dati includono i voti degli studenti, caratteristiche demografiche, sociali e scolastiche ed è stato raccolto utilizzando registri scolastici e questionari. Alcuni degli attributi più rilevanti sono riportati di seguito:

- School: GP o MS in base alla scuola i provenienza
- Sex: sesso anagrafico dello studente (F o M)
- Address: contesto urbano in cui vive lo studente urbano (U) o rurale (R)
- Medu: livello di educazione della madre (da 0 che indica nessuno a 4 che indica istruzione superiore)
- Fedu: livello di educazione del padre (stessa scala di Fedu)
- Studytime: tempo di studio settimanale (da 1 che indica meno di due ore a 4 che indica più di 10 ore)
- Failures: numero di voti finali insufficienti negli anni precedenti (da 0 a 4)
- Higher: vuole proseguire con gli studi superiori (sì o no)
- Internet: ha accesso a internet a casa (sì o no)
- Walc: consumo alcolico nei giorni lavorativi (da 1 che indica molto basso a 5 molto alto)
- Absences: indica il numero di assenze dello studente (da 0 a 93)
- G1, G2, G3: voto finale del primo, secondo e terzo periodo in cui è diviso l'anno scolastico (da 0 a 20 in cui la sufficienza è 10)

Qua di seguito sono riportate le prime 10 righe del del dataset:

school	sex	age	adc	Medu	Fedu	Mjob	Fjob	studytime	failures	higher	internet	Dalc	Walc	absences	G1	G2	G3
GP	F	18	U	4	4	at_hom	teacher	2	0	yes	no	1	1	4	0	11	11
GP	F	17	U	1	1	at_hom	other	2	0	yes	yes	1	1	2	9	11	11
GP	F	15	U	1	1	at_hom	other	2	0	yes	yes	2	3	6	12	13	12
GP	F	15	U	4	2	health	service	3	0	yes	yes	1	1	0	14	14	14
GP	F	16	U	3	3	other	other	2	0	yes	no	1	2	0	11	13	13
GP	M	16	U	4	3	service	other	2	0	yes	yes	1	2	6	12	12	13
GP	M	16	U	2	2	other	other	2	0	yes	yes	1	1	0	13	12	13
GP	F	17	U	4	4	other	teacher	2	0	yes	no	1	1	2	10	13	13
GP	M	15	U	3	2	service	other	2	0	yes	yes	1	1	0	15	16	17
GP	M	15	U	3	4	other	other	2	0	yes	yes	1	1	0	12	12	13

[3] https://archive.ics.uci.edu/ml/datasets/Student+Performance

[4] https://colab.research.google.com/drive/18UyaS9Sio2F_KdQobJg0PGkG8yNrFsXA#scrollTo=4iUBUAZ4JrT9

Preprocessing del dataset

La fase di preprocessing di questo esperimento è stata simile a quella precedente. Sono stati rimossi gli attributi meno significativi con valori molto generici, ma anche gli attributi riguardanti i voti del primo, secondo e terzo periodo dell'anno scolastico (G1, G2 e G3). La rimozione di questi attributi, relativi ai voti degli studenti, rende più difficile predire il valore dell'attributo target per l'algoritmo, ma la predizione fatta in assenza di questi, è più utile perché si basa molto di più su tutte le restanti caratteristiche dello studente. Infatti se non venissero rimossi, l'algoritmo utilizzerebbe quasi unicamente questi attributi trascurando gli altri e rendendo così l'individuazione di possibili bias difficile. Tuttavia è comunque necessario avere un attributo che faccia riferimento alla performance degli studenti, e per questo è stato aggiunto "G_avg" che rappresenta la media degli attributi rimossi G1, G2 e G3.

Sono stati poi scelti gli attributi per il RandomForestClassifier, algoritmo che abbiamo già utilizzato per il dataset precedente. In particolare "G_avg" è stato scelto come attributo target che l'algoritmo deve predire, "studytime" è stato scelto come attributo legittimo e "address", che ricordiamo indica se lo studente vive in un contesto urbano o rurale, è stato scelto come attributo sensibile. I valori dell'attributo target e dell'attributo legittimo sono stati mappati in modo che fossero binari. L'attributo "G_avg" è stato mappato in 0 e 1 a seconda che lo studente avesse un valore di media dei voti sopra o sotto la sufficienza, così come l'attributo studytime è stato mappato in 0 e 1 a seconda che il tempo di studio settimanale del singolo studente fosse sopra o sotto la media degli altri studenti.

Infine gli studenti sono stati divisi nei gruppi discriminato e privilegiato in base ai valori dell'attributo "address", gli studenti rural sono stati considerati parte del gruppo discriminato mentre quelli urban del gruppo privilegiato.

```
#save the two groups indexes
Y = df[target_variable]
X = df.drop(target_variable, axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=1
#y_true=df[target_variable].loc[list(X_test.index)]
sensible_indexes=df[sensible_attribute].loc[list(X_test.index)]
legittimate_indexes=df[legittimate_attribute].loc[list(X_test.index)]
```

```
#save the two groups indexes
discriminated = []
privileged = []
for idx, i in enumerate(sensible_indexes):
  if i==sensible_values[0]:
    discriminated.append(sensible_indexes.index[idx])
  else:
    privileged.append(sensible_indexes.index[idx])
```

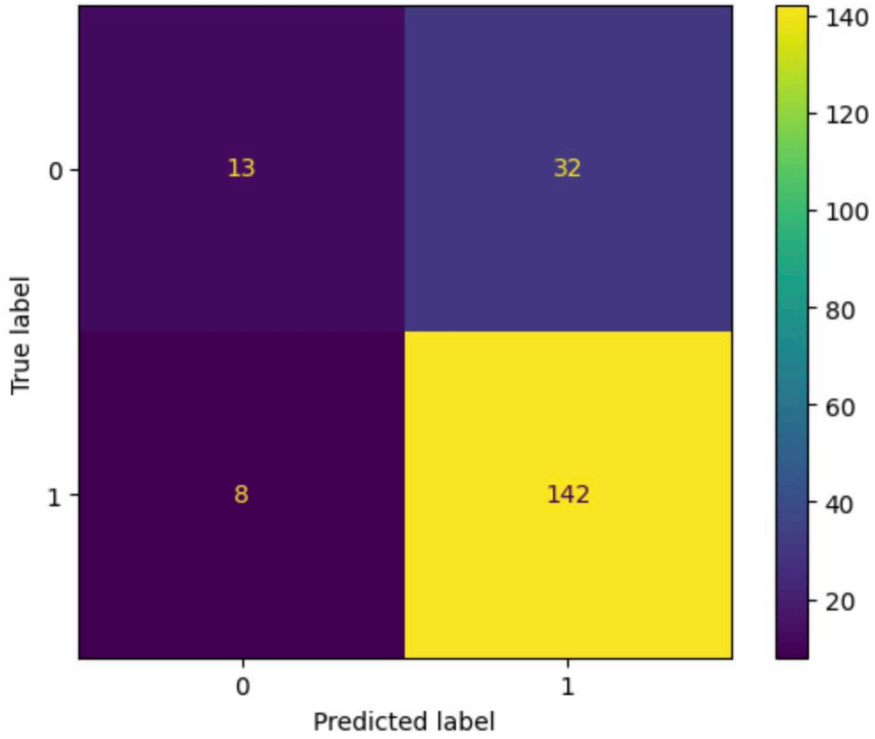
Codice per dividere gli studenti nel gruppo privilegiato e discriminato in base al valore dell'attributo sensibile ("address")

Processing del dataset

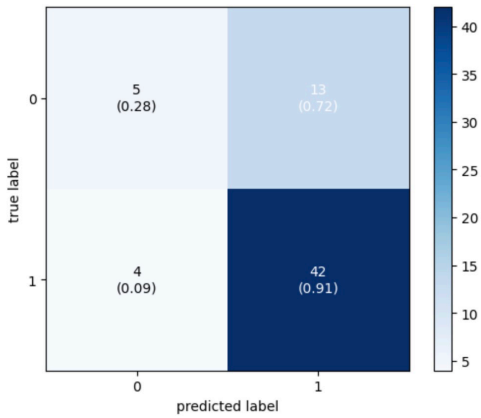
Le fasi del processing sono le stesse dell'analisi precedente. Il train set include anche in questo caso il 70% (454 studenti) degli studenti totali mentre il test set la restante parte (195).

Esiti del Processing

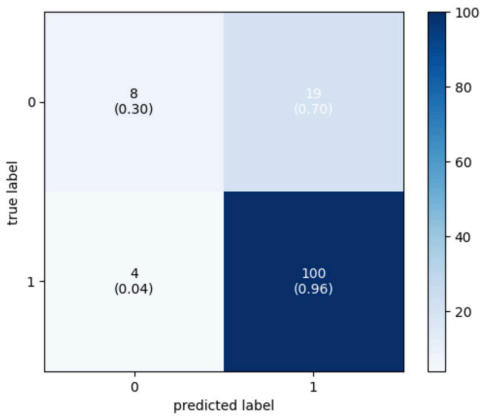
La confusion matrix che rappresenta i 4 gruppi del test diagnostico individuati a valle della fase di processing è la seguente:



L'accuratezza dell'algoritmo è stata del 79.48% ed è stata ricavata sempre dalla formula (TP+TN)/(TP+TN+FP+FN), cioè il rapporto degli studenti correttamente individuati (True Positive + True Negative) e del numero totale degli studenti del test set di cui l'algoritmo ha predetto il valore di "G_avg". Tale accuratezza ci soddisfa e ci permette di proseguire con l'analisi del dataset. Vengono quindi analizzate le confusion Matrix dei singoli gruppi in modo da rilevare la presenza eventuale di bias e unfairness. Le due confusion Matrix sono le seguenti:



Discriminato



Privilegiato

Fairness Metrics Computation

Una volta ottenute le due confusion Matrix relative al gruppo discriminato e privilegiato siamo pronti a calcolare i valori delle diverse metriche che abbiamo già utilizzato per l'analisi del dataset precedente in modo da individuare la presenza di bias o unfairness nel dataset. Il calcolo delle metriche ha portato ai seguenti risultati:

	Metric	Value	Value_discriminated_group	Value_privileged_group
0	GroupFairness	0.946035	0.859375	0.908397
1	ConditionalStatisticalParity	0.993883	0.925000	0.930693
2	PredictiveParity	0.908727	0.763636	0.840336
3	PredictiveEquality	1.026316	0.722222	0.703704
4	EqualOpportunity	0.442308	0.086957	0.038462
5	EqualizedOdds	0.974554	0.949565	1.026316
6	ConditionalUseAccuracyEquality	0.757273	0.908727	0.833333
7	OverallAccuracyEquality	0.262500	0.420000	0.625000
8	TreatmentEquality	1.400493	19.692308	27.578947
9	FORParity	0.750000	0.444444	0.333333
10	FN	0.488550	0.062500	0.030534
11	FP	1.400493	0.203125	0.145038

Come già fatto nell'analisi del dataset precedente scegliamo l'intervallo [0.9, 1,1] entro cui considerare i valori delle metriche come accettabili.

In particolare osservando i valori delle varie metriche ci accorgiamo che quelle che non rientrano nello spettro di accettazione escludendo FN e FP dato che dipendono fortemente dalla dimensione dei singoli gruppi sono:

- Equal Opportunity
- Condintional Use Accuracy Equality
- Overall Accuracy Equality
- Treatment Equality
- FOR Parity

L'Equal Opportunity presenta una differenza di 0.66 rispetto al valore ideale (1), e questo indica che gli studenti che vivono in un contesto rurale hanno una possibilità maggiore di essere predetti avere una media dei voti finali insufficiente nonostante abbiano in realtà avuto una media sufficiente. Conditional Use Accuracy Equality si discosta di 0.24 rispetto al valore ideale e significa che gli studenti del gruppo discriminato che sono effettivamente positivi (hanno una media sufficiente) hanno una minore possibilità di essere predetti come positivi, e gli studenti di questo stesso gruppo che sono effettivamente negativi hanno una minore possibilità di essere predetti come negativi rispetto agli studenti dell'altro gruppo. Questo indica che c'è stata un'accuratezza minore nei confronti del gruppo discriminato. L'Overall Accuracy Equality è di 0.74 inferiore rispetto al valore ideale, questa metrica mette a confronto il numero di veri positivi e veri negativi individuati nei due gruppi a prescindere dal numero di FP e FN degli stessi. In questo caso come nella metrica precedente il valore della metrica ci indica che l'accuratezza dell'algoritmo è stata inferiore nel caso del gruppo discriminato. Nonostante ciò essendoci una grande differenza nella numerosità dei due gruppi (64 e 131) in questo caso è più opportuno fare affidamento al

valore della metrica precedente in quanto per ogni gruppo tiene conto del numero percentuale (e non assoluto come in questo caso) di TN e TP rispetto al numero totale di Negativi e Positivi individuati. Treatment equality considera il rapporto tra i FN e FP di ogni gruppo moltiplicandolo poi per la dimensione del gruppo corrispondente. In questo caso il valore della metrica è molto sbilanciato in quanto essendoci un fattore moltiplicativo per la lunghezza, ed essendo il gruppo privilegiato molto maggiore. Considerando unicamente i rapporti tra FN e FP osserviamo che a parità di decisione errata per il gruppo discriminato è stato più difficile che l'algoritmo prevedesse un valore positivo rispetto agli studenti del gruppo privilegiato. Il valore di FOR Parity indica che gli individui del gruppo discriminato, che in base alle loro caratteristiche dovrebbero ricevere una previsione negativa, hanno una probabilità più alta di ricevere una previsione positiva rispetto agli studenti del gruppo privilegiato che allo stesso modo dovrebbero ricevere una previsione negativa in base alle loro caratteristiche. Questa metrica è strettamente legata all'accuratezza inferiore dell'algoritmo nei confronti del gruppo discriminato.

Dopo aver analizzato singolarmente le metriche ed aver compreso la motivazione per cui alcuni di questi valori possono essere in contrasto con gli altri, possiamo concludere che il gruppo di studenti che popolano questo dataset che vivono in un contesto urbano siano più avvantaggiati rispetto a quelli che vivono in un contesto rurale.

Conclusioni

Per eseguire questa analisi è stato fatto uno studio preliminare riguardo al significato dell'etica e delle sue diverse componenti nell'ambito informatico, utilizzando diversi articoli presenti in letteratura [9][10]. In particolare è stato approfondito il concetto della fairness nei dati e le sue implicazioni, per poi applicare questi principi ai due dataset che abbiamo studiato. La scelta delle metriche per la valutazione della fairness è stata fatta in modo da comprendere le metriche che considerassero principalmente gli aspetti di Group fairness, in quanto il nostro studio si è basato sull'individuazione di unfairness o pregiudizi sistemici che possono verificarsi a livello di gruppo, tra queste per esempio troviamo la metrica di Group fairness già analizzata in precedenza.

Dopo aver selezionato le metriche adeguate, si è deciso di analizzare due dataset che riguardano un argomento studiato meno frequentemente nell'ambito della fairness: l'educazione. In questo contesto, si è meno portati a considerare la presenza di disuguaglianze sistematiche rispetto ad altri campi. Sono stati scelti quindi due dataset che contengono informazioni riguardanti gli studenti, le loro prestazioni scolastiche, e i risultati in termini di media e punteggi ai test. Manipolando i dati di ciascun dataset e calcolando i valori delle metriche adottate per questa ricerca, abbiamo analizzato i dati e riscontrato la presenza di bias in entrambi i dataset.

Il fatto che siano presenti dei bias implica che i risultati scolastici che uno studente ottiene non derivano unicamente dalle sue capacità, ma sono basati anche su caratteristiche che in realtà non dovrebbero competere nel determinare i risultati didattici dello stesso. In particolare nel primo dataset il fato che siano presenti dei bias significa che uno studente di sesso maschile ha conseguito un punteggio maggiore nel test SAT rispetto a uno studente di sesso femminile a

parità di media scolastica dei due. Nel secondo caso invece, il fatto che siano presenti dei bias, implica che a parità di tempo di studi, e di condizioni che comprendono la vita sociale come il consumo di alcolici e la situazione familiare, uno studente che vive in un contesto urbano ha un rendimento scolastico migliore rispetto ad uno studente che vive in un contesto rurale.

È chiaro che la scelta dell'attributo sensibile è significativa, proprio perché nel caso in cui vengano rilevati dei bias nei dati, si riconduce a questo attributo la causa della discriminazione. Dunque bisogna prestare particolare attenzione a scegliere un attributo sul quale la decisione non dovrebbe basarsi, come il sesso anagrafico o la zona di residenza, e non su attributi che concorrono in modo giustificato nelle differenze tra gli individui, come per esempio il tempo di studio.

I due casi analizzati sono solo alcuni esempi nel campo dell'apprendimento che presentano dei bias, e l'individuazione di quest'ultimi ci aiuta a eliminare le disparità esistenti e a prevenirne di nuove, creando un ambiente più sano e con possibilità di crescita uguali per tutti. Essendo l'educazione alla base della società, formare le persone in un contesto giusto fa si che ci siano più possibilità che le disparità presenti in altri contesti sociali diminuiscano.

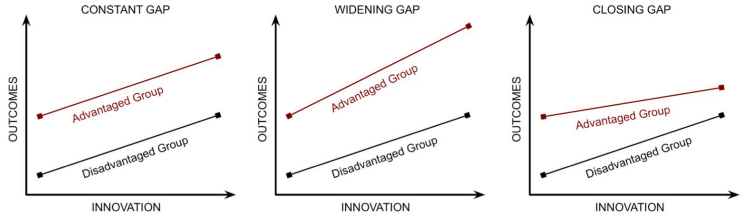
È bene notare che queste tipologie di studi sono ancora in fase iniziale per quanto riguarda l'ambito educativo. Tuttavia sono importanti poiché pongono le basi per identificare i bias già presenti nel sistema scolastico e per prevenire l'emergere di nuovi. Infatti grazie alla tecnologia, stiamo assistendo a una rivoluzione nell'ambito pedagogico, che ha riportato l'attenzione su questioni cruciali in merito alla giustizia e all'equità.

In particolare l'istruzione sta diventando sempre più accessibile, collaborativa e su misura, attraverso l'utilizzo di strumenti che permettono di personalizzare l'apprendimento, formarsi online e analizzare le performance degli studenti.

Tuttavia l'impiego di algoritmi che modificano l'esperienza di apprendimento, e che valutano gli studenti, richiede grande attenzione per evitare che le disuguaglianze già presenti vengano accentuate e che ne nascano di nuove.

Il grafico qua di seguito rappresenta tre possibili andamenti della differenza di trattamento tra un gruppo avvantaggiato, e uno svantaggiato, in seguito all'introduzione di innovazioni. Nel primo caso, nonostante l'innovazione (rappresentata dall'evoluzione del sistema educativo attraverso la tecnologia), la differenza di trattamento rimane invariata. Nel secondo scenario, i cambiamenti introdotti accentuano le disuguaglianze tra i due gruppi, ed è ciò che si vuole evitare studiando la fairness delle novità che vengono introdotte. L'ultimo scenario, invece, è quello auspicabile ed è quello in cui l'innovazione comporta maggiori benefici, non solo aumentando il benessere di chi utilizza l'infrastruttura che viene rinnovata, ma anche diminuendo la disparità di trattamento tra gli individui.

Pertanto, è fondamentale disporre di strumenti efficaci per valutare gli algoritmi e garantire che vengano utilizzati solo per migliorare l'ambito in cui vengono impiegati, che in questo caso è quello dell'educazione.



[8]

[1] Terrel Ward Bynum. 2000. A very short history of computer ethics. https://www.cs.utexas.edu/~ear/cs349/Bynum_Short_History.html

[2] June Iqbal e Bilal Maqbool Beigh. 2017. https://www.researchgate.net/publication/318452200_Computer_Ethics_from_Obscure_to_Ubiquitous

[3] James Moor. 1985. What is computer ethics?. https://web.cs.ucdavis.edu/~rogaway/classes/188/spring06/papers/moor.html

[4] Letizia Tanca, Donatella Firmani e Riccardo Torlone. 2021. Etica e qualità dei dati: metodi e strumenti per rendere gli algoritmi responsabili. https://www.agendadigitale.eu/industry-4-0/etica-dei-dati-by-design-metodi-e-strumenti-per-evitare-le-disuguaglianze-degli-algoritmi/

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

[6] Sahil Verma e Julia Rubin. 2018. Fairness Definitions Explained. https://fairware.cs.umass.edu/papers/Verma.pdf

[7] Algoritmo random forest classification. https://www.ibm.com/it-it/topics/random-forest

[8] René Kizilcec e Hansol Lee. Algorithmic Fairness in Education. https://arxiv.org/pdf/2007.05443.pdf

[9] Julia Stoyanovich, Serge Abiteboul e Gerome Miklau. Data, Responsibly:
Fairness, Neutrality and Transparency in Data Analysis. https://openproceedings.org/2016/conf/edbt/paper-c.pdf

[10] H. V. Jagadish, Julia Stoyanovich e Bill Howe. The Many Facets of Data Equity. https://ceur-ws.org/Vol-2841/PIE+Q_6.pdf