UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Sostenibilità Sociale: uno Studio Empirico sui Costi degli Strumenti di Fairness

RELATORE

Prof. Fabio Palomba

Dott. Vincenzo De Martino

CANDIDATO

Ciro Troiano

Matricola: 0512111005

Anno Accademico 2022-2023

Questa tesi è stata realizzata nel

"Don't let yourself make excuses for not doing the things you want to do."

Sam Altman

Questa tesi ha contribuito a piantare un albero in Madagascar tramite il progetto Treedom.

https://www.treedom.net/it/user/sesalab/event/sesa-random-forest

Abstract

Al giorno d'oggi l'Intelligenza Artificiale è al centro dello sviluppo tecnologico e informatico e la tendenza di ogni settore è puntare a soluzioni e servizi che utilizzino sistemi AI. Il settore dell'Intelligenza Artificiale non è un settore "nuovo", sistemi che utilizzavano soluzioni che ad oggi chiameremo "AI" esistono da molto prima dell'avvento dei sistemi moderni che rappresentano vere e proprie rivoluzioni dell'evoluzione tecnologica.

In particolare, realizzare un modello di Machine Learning rappresenta da sempre un processo fondamentale per la creazione di un sistema AI. Realizzare un "buon" modello di Machine Learning non è semplice e bisogna tenere a mente diverse problematiche legate non solo all'utilizzo dei dati, in particolare il concetto di Fairness, ma anche alla Quality del prodotto finale e ai costi sostenuti per lo sviluppo e la manutenzione.

Questo studio nasce dalla mancanza di articoli scientifici importanti che analizzino i trade-offs sostenuti per ottenere un prodotto finito di qualità che rispetti anche condizioni di equità nel trattamento delle informazioni. In particolare, questo studio di benchmark, sfruttando delle librerie di Fairness e di Sostenibilità presenti sul mercato, andrà ad esaminare tutte queste situazioni, evidenziando singolarmente i trade-offs necessari, incentivando o sconsigliando, l'uso di possibili tecniche per ottenere dei prodotti finiti che rispettino standard non solo di Qualità ma di Equità e Sostenibilità del prodotto e del processo di produzione stesso.

Indice

Elenco delle Figure **iii**

Elenco delle Tabelle **iv**

1 Introduzione **1**

1.1 Contesto 1

1.2 Motivazioni e obiettivi 2

1.3 Risultati 3

1.4 Struttura della tesi 4

2 Stato dell'arte **5**

2.1 L'Intelligenza Artificiale e il Machine Learning 5

2.1.1 Il Machine Learning 5

2.2 Machine Learning Quality 7

2.3 Machine Learning Fairness 10

2.3.1 Strategie per costruire un modello fair di ML 14

2.4 Sustainability 15

3 Metodologia **18**

3.1 Motivazioni e obiettivi dello studio 18

3.2 Variabili indipendenti dello studio 21

3.2.1 I datasets 21
3.2.2 I gruppi protetti e non protetti 24
3.2.3 Gli algoritmi di ML e le reti neurali di DL scelte 26
3.2.4 I tool di Fairness utilizzati 27
3.2.5 Il tool di Sustainability utilizzato 30
3.2.6 L'ambiente di lavoro 31
3.2.7 La macchina di esecuzione 31
3.3 Variabili dipendenti dello studio 32
3.3.1 Le metriche di equità 32
3.3.2 Le metriche di sostenibilità 34
3.3.3 Le metriche qualitative 35
3.3.4 Training e Testing dei modelli 36
3.3.5 L'analisi dei risultati 37
3.4 La struttura del progetto 38

4 Analisi dei risultati **40**
4.1 RQ1: Che impatto hanno gli strumenti di Fairness sulla sostenibilità sociale del prodotto? 41
4.2 RQ2: Che impatto hanno gli strumenti di Fairness sulla sostenibilità ambientale del prodotto? 55
4.3 RQ3: Che impatto hanno gli strumenti di Fairness sulla sostenibilità economica del prodotto? 61

5 Conclusioni **79**

Bibliografia **81**

Elenco delle figure

2.1 Il processo iterativo dell'apprendimento supervisionato. 6
2.2 Il processo di creazione di un algoritmo di apprendimento supervisionato e il suo funzionamento. 7
2.3 Il modello IXI per lo sviluppo di sistemi ML. 9
2.4 Esempi di bias nelle varie interazioni fra algoritmo, utente e dati. . . 10
2.5 La fair pipeline definita da Bellamy et al. utilizzata per realizzare un modello fair di Machine Learning. 14

3.1 La pipeline di produzione dell'intero studio. 39

Elenco delle tabelle

3.1 I dataset scelti e le loro caratteristiche 25
3.2 I dataset scelti e alcune caratteristiche sui gruppi protetti e non protetti 25

4.1 Le metriche di Fairness calcolate prima e dopo le operazioni di preprocessing sull'Adult Dataset. 42
4.2 Le metriche di Fairness calcolate prima e dopo le operazioni di preprocessing sul German Credit Dataset. 43
4.3 Le metriche di Fairness calcolate prima e dopo le operazioni di preprocessing sul Heart Disease Dataset. 44
4.4 Le metriche di Fairness calcolate prima e dopo le operazioni di preprocessing sul Home Credit Default Risk Dataset. 45
4.5 Le metriche di Fairness calcolate prima e dopo le operazioni di preprocessing sull'UTKFace Dataset (DL). 46
4.6 Le metriche di Fairness calcolate prima e dopo le operazioni di inprocessing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 47
4.7 Le metriche di Fairness calcolate prima e dopo le operazioni di inprocessing sull'UTKFace Dataset (DL). 47

4.8 Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 48
4.9 Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 48
4.10 Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 49
4.11 Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 50
4.12 Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 51
4.13 Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 52
4.14 Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact) 53
4.15 Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sull'UTKFace Dataset (DL). 53
4.16 I consumi dei modelli standard di ML sui diversi datasets. 56
4.17 I consumi dei modelli standard di DL sull'UTKFace dataset. 56

4.18 I consumi dei modelli di ML, in ordine di operazione di pre-in-post-processing, per i vari datasets (AIF360). 57
4.19 I consumi dei modelli di DL, in ordine di di pre-in-post-processing, sull'UTKFace dataset (AIF360). (RN50 = ResNet50, MNV2 = MobileNetV2) 57
4.20 I consumi dei modelli di ML, in ordine di operazione di pre-in-post-processing, per i vari datasets (Fairlearn). 58
4.21 I consumi dei modelli di DL, in ordine di di pre-in-post-processing, sull'UTKFace dataset (Fairlearn). (RN50 = ResNet50, MNV2 = MobileNetV2) 58
4.22 Le metriche ottenute dai modelli standard sull'Adult Dataset. 61
4.23 Le metriche ottenute dai modelli standard sul German Credit Dataset. 61
4.24 Le metriche ottenute dai modelli standard sul Heart Disease Dataset. 62
4.25 Le metriche ottenute dai modelli standard sul Home Credit Default Risk Dataset. 62
4.26 Le metriche di qualità ottenute dai modelli di DL sull'UTKFace Dataset. 62
4.27 Le metriche ottenute dai modelli con pre-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) . . 64
4.28 Le metriche ottenute dai modelli con pre-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn 64
4.29 Le metriche ottenute dai modelli con pre-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn 65
4.30 Le metriche ottenute dai modelli con pre-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn 65

4.31 Le metriche di qualità ottenute dai modelli di DL con pre-processing sull'UTKFace Dataset. (AIF = AIFairness360, FLN = Fairlearn) 66
4.32 Le metriche ottenute dai modelli con in-processing sull'Adult Dataset. 67
4.33 Le metriche ottenute dai modelli con in-processing sul German Credit Dataset. 67
4.34 Le metriche ottenute dai modelli con in-processing sul Heart Disease Dataset. 67
4.35 Le metriche ottenute dai modelli con in-processing sul Home Credit Default Risk Dataset. 68
4.36 Le metriche di qualità ottenute dai modelli di DL con in-processing sull'UTKFace Dataset. 68
4.37 Le metriche ottenute dai modelli con post-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 69
4.38 Le metriche ottenute dai modelli con post-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 70
4.39 Le metriche ottenute dai modelli con post-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 70
4.40 Le metriche ottenute dai modelli con post-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 71
4.41 Le metriche di qualità ottenute dai modelli di DL con post-processing sull'UTKFace Dataset. (AIF = AIFairness360, FLN = Fairlearn) 71
4.42 I tempi medi necessari ad eseguire i vari script con pre-processing (AIF360-Fairlearn) sui diversi datasets (in secondi). 73
4.43 I tempi medi necessari ad eseguire i vari script con in-processing (AIF360) sui diversi datasets (in secondi). 73

4.44 I tempi necessari ad eseguire i vari script con post-processing (AIF360-Fairlearn) sui diversi datasets (in secondi). 74
4.45 I tempi necessari ad eseguire i vari script dei modelli DL con pre-processing (AIF360-Fairlearn) sull'UTKFace Dataset (in secondi). . . 74
4.46 I tempi necessari ad eseguire i vari script dei modelli DL con in-processing e post-processing (AIF360) sull'UTKFace Dataset (in secondi). 74
4.47 I pesi (in MB) ottenuti dai vari modelli sull'Adult Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 75
4.48 I pesi (in MB) ottenuti dai vari modelli sul German Credit Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) . . 76
4.49 I pesi (in MB) ottenuti dai vari modelli sul Heart Disease Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) . . 76
4.50 I pesi (in MB) ottenuti dai vari modelli sul Home Credit Default Risk Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn) 77
4.51 I pesi (in MB) ottenuti dai vari modelli sull'UTKFace Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing. (AIF = AIFairness360, FLN = Fairlearn, RN50 = ResNet50, MNV2 = MobileNetV2) 77

CAPITOLO 1

Introduzione

1.1 Contesto

L'Intelligenza Artificiale è diventata fulcro dello sviluppo informatico e tecnologico [1]. Ogni settore, ad oggi, presenta delle soluzioni e servizi che usano interamente, o in parte, un sistema AI [2]. Lo sviluppo di un sistema AI si basa sulla disponibilità costante di dati. Bisogna ricordare che alla base di un Intelligenza Artificiale ci sono quantità innumerevoli di dati sulla quale il sistema basa tutta la sua "conoscenza" [3]. I dati però, inevitabilmente, presentano Bias, ovvero una classe di errori sistematici legati ai singoli dati, alla loro rappresentazione, al modo in cui sono stati raccolti ed alle informazioni che rappresentano, che possono influire, anche in maniera esponenziale, non solo sulla qualità del prodotto ma in particolare anche sull'etica delle "scelte" del prodotto finale. Non sono rari, infatti, casi di sistemi IA che presentano delle forti discriminazioni di tipo razziale, *gender-based* e legate alla forte discrepanza nei dati [4, 5].

Alla base dei sistemi AI ci sono modelli di Machine Learning e di Deep Learning e in questo studio ci concentreremo fondamentalmente sui modelli di Machine Learning, con qualche primo approccio ai modelli di Deep Learning, andando ad esaminare come le pratiche più comuni di Fairness possano impattare sul risultato e sulla qualità del prodotto finale e sui costi, ambientali e aziendali, sostenuti per realizzare e manutenere un modello conforme agli standard di Fairness e che presenti una qualità del prodotto non indifferente.

1.2 Motivazioni e obiettivi

L'idea di questo lavoro nasce dall'interesse personale nell'ambito AI, uno degli ambiti più popolari al giorno d'oggi. In particolare, sono sempre stato affascinato sia dal successo di grandi progetti AI che rivoluzionano la tecnologia attuale ma ancora di più ai "fallimenti" più eclatanti del settore e dalle motivazioni legate ad essi [2].
I risultati sono da sempre importanti in questo contesto ma con l'interesse costante che il settore sta ricevendo anche il comportamento oggettivo ed equo dei sistemi sta diventando fulcro di discussione, in particolare parliamo del concetto di equità, o di Fairness, dei sistemi AI.
Anche se il concetto di Fairness è presente sin dagli albori nei settori della Software Engineering e dell'AI, con l'evoluzione sempre costante dei sistemi AI sia in complessità che in termini degli obbiettivi proposti, è diventato uno dei punti cardine della praticità di una soluzione AI, ad oggi, infatti, fornire dei risultati oggettivamente corretti e privi di discriminazione è diventato importante tanto quanto fornire delle soluzioni che presentano elevata accuratezza nel modo in cui operano [6].
Un altro tema che sta ricevendo sempre più attenzione mediatica in AI è la sostenibilità, purtroppo quando vengono presentate delle nuove soluzioni ground-breaking nel settore si mettono in secondo piano molti aspetti "secondari" della creazione e sviluppo del sistema, uno di questi temi è proprio la sostenibilità del processo e del sistema stesso, non è raro infatti che le soluzioni più importanti in ambito AI abbiano richiesto un dispendio energetico e di risorse talmente imponente che spesso richiedevano dovuti trade-offs delle risorse richieste e della qualità del sistema.

Sebbene questi due argomenti siano sempre più rilevanti nel contesto, in letteratura, sono pochissimi gli studi che esaminano i trade-off richiesti per ottenere modelli equi nel rispetto delle risorse e della qualità del prodotto realizzato.
Questa mancanza in letteratura è proprio uno dei punti cardine di questo studio, infatti, l'obiettivo principale è studiare e valutare come le tecniche di Fairness influenzino la qualità e la sostenibilità dei prodotti realizzati, andando ad esaminare i vari trade-offs da sempre evidenziati ma che, spesso, hanno trovato poco spazio di approfondimento e di analisi, tramite uno studio di benchmarking, sfruttando due librerie di Fairness disponibili sul mercato.

1.3 Risultati

I risultati di questo studio hanno permesso di avere una visione molto chiara dei *trade-offs* incontrati durante lo sviluppo di un modello per ottenere dei risultati "equi" nel modo in cui classifica e predice i dati forniti. Dai risultati ottenuti è emerso come, spesso, creare modelli "equi" trattando in diverse fasi i dati o i modelli nelle diverse fasi della creazione di un modello, risulti essere abbastanza ininfluente, in termini di qualità, permettendo quindi di creare modelli di qualità ma molto più "equi" nel loro comportamento. Inoltre, è stato possibile evidenziare come i modelli, in particolare modelli realizzati con DL, abbiano dei costi non indifferenti e che operazioni di pre-processing, in-processing e post-processing sui dati e sulle predizioni del modello possano ulteriormente appesantire il tempo e le risorse richieste, aumentando esponenzialmente i consumi richiesti.

1.4 Struttura della tesi

Il lavoro di tesi è suddiviso nei seguenti capitoli:

- **Capitolo 2: Stato dell'arte**, in questo capitolo viene presentata un'indagine individuale sui concetti di Machine Learning, Fairness, Sustainability e Quality in letteratura.
- **Capitolo 3: Metodologia e ricerca**, in questo capitolo viene descritto formalmente l'obiettivo principale dello studio, vengono presentati i quesiti di ricerca formulati e gli strumenti utilizzati per fornire una risposta concreta.
- **Capitolo 4: Analisi dei risultati**, in questo capitolo vengono esposti i risultati dell'analisi su diversi modelli di ML e di DL, utili per poter fornire delle riposte alle domande di ricerca presentate.
- **Capitolo 5: Conclusioni**, viene effettuata una sintesi del lavoro svolto e le possibili applicazioni per lavori futuri.

CAPITOLO 2

Stato dell'arte

2.1 L'Intelligenza Artificiale e il Machine Learning

2.1.1 Il Machine Learning

Quando si parla di Intelligenza Artificiale, si considera formalmente il concetto di una macchina in grado di emulare il comportamento umano [7, 8, 9, 10].
Il **Machine Learning (ML)** è una branca dell'Informatica che ha radici in una molteplicità di discipline diverse come la statistica, la complessità computazionale ma anche la filosofia, biologia e scienza cognitiva e si occupa, formalmente, di sviluppare algoritmi e modelli di apprendimento automatico che consentono al software di migliorare automaticamente la sua precisione e le sue capacità attraverso i dati e l'esperienza accumulata nel tempo [11, 12].
Ad oggi il Machine Learning è alla base di numerose applicazioni come il riconoscimento delle immagini, valutazione di rischi, stime, predizioni e strategie di marketing.
Per Machine Learning si intendono algoritmi che sfruttano un'enorme quantità di dati, definiti tradizionalmente come "dataset", per poter permettere all'algoritmo di determinare pattern e caratteristiche nei dati che possono permettere a quest'ultimo di prendere decisioni e ottenere risultati.

Per realizzare un modello viene utilizzato un processo di apprendimento e valutazione dei risultati di tipo iterativo, infatti ripetendo più volte queste operazioni ci permette di ottenere un modello in grado di fornire degli output quanto più realistici possibili [8, 13].

Gli algoritmi di ML vengono tradizionalmente classificati in base al tipo di output desiderato, in *apprendimento supervisionato* e *apprendimento non supervisionato* [13], che si differenziano sostanzialmente nella presenza o meno di "etichette" poste sulle informazioni che l'algoritmo sfrutta per individuare dei pattern e caratteristiche sulle istanze di input. Nell'apprendimento supervisionato, Osisanwo et al. [14] definiscono il processo iterativo dell'apprendimento supervisionato in **Figura 2.1**.

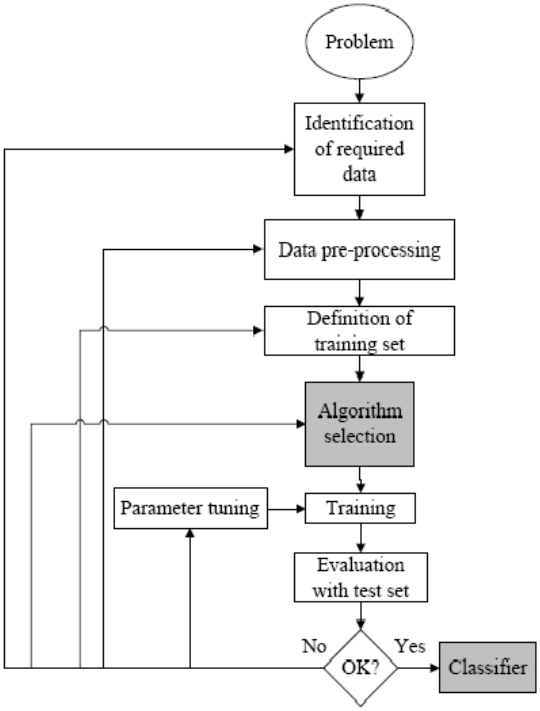

Figura 2.1: Il processo iterativo dell'apprendimento supervisionato.

In generale, V. Nateski [13] fornisce una visione globale del processo di creazione di un algoritmo di apprendimento supervisionato e il suo funzionamento in **Figura 2.2**.
Dai riferimenti [13, 14, 15, 16] è possibile approfondire la classificazione e descrizione dettagliata degli algoritmi di apprendimento supervisionato più importanti.

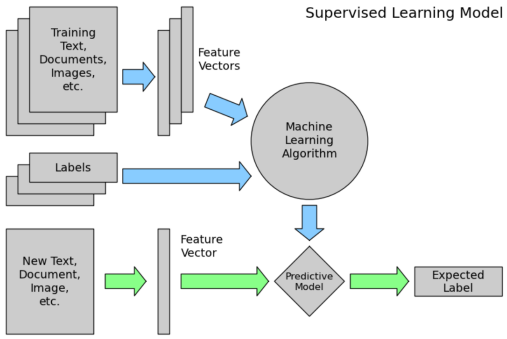

Figura 2.2: Il processo di creazione di un algoritmo di apprendimento supervisionato e il suo funzionamento.

2.2 Machine Learning Quality

Come tutti i progetti software degli ultimi decenni, particolare importanza viene data alla qualità dell'artefatto software realizzato. Nell'Intelligenza Artificiale, e in particolare nel Machine Learning, realizzare un prodotto finale che possa corrispondere a determinati standard di qualità è un requisito critico per la realizzazione di un prodotto software di buona qualità.
Sebbene in Software Engineering (SE), la qualità di un prodotto sia ampiamente discussa e valutata, solamente negli ultimi anni questo contesto di qualità si è diffuso anche nel mondo dell'intelligenza artificiale [17].
Formalmente il concetto di qualità in ambito AI viene ripreso dal concetto di qualità del software ma è chiaro come le sfide affrontate per realizzare un modello di ML ed un sistema AI siano sostanzialmente diverse, o case-specific, rispetto ad un sistema software in SE [17, 18, 19].
Uno degli ultimi utilizzi più estensivi dell'AI è nel campo della Software Engineering, nel quale sono state definite delle nuove strategie che sfruttano i nuovi sistemi AI per

sostenere lo sviluppo software, strategia che prende il nome di *AI4SE* [20, 21]. Negli ultimi anni, invece, è stata data particolare attenzione all'utilizzo delle strategie SE per realizzare dei sistemi AI, che prende il nome di *SE4AI*, in contrapposizione alla strategia precedente [21].

Il concetto di *SE4AI* è ancora agli albori, infatti, dallo studio condotto da S. Masuda et al. [22] si evince come il numero di articoli in merito sia relativamente basso in contrapposizione agli articoli su concetti e argomenti di rilevanza, ma da uno studio successivo di S. Martinez-Fernàndez et al. si evidenzia un particolare interesse al contesto che cresce di anno in anno, coinvolgendo diverse realtà industriali, soprattutto in Europa e Nord America [17].

Il concetto di *SE4AI* nasce dalle necessità di fornire delle chiare direttive sullo sviluppo di sistemi AI, di cui oggi è possibile vedere un costante aumento di richiesta in ogni applicazione industriale.

Il Machine Learning tende, per natura, a sviluppare dei modelli black-box *fortemente dipendenti* da metriche e statistiche che spesso rendono particolarmente difficile descrivere il comportamento di quest'ultimi, infatti, circa il 40% degli ingegneri software dediti allo sviluppo AI considera difficile fornire un sistema AI di qualità [23]. Da uno studio condotto nel 2019 da S. Amershi et al. [24] è stato evidenziato come tutte le aree di un comune processo di sviluppo di un modello di ML siano direttamente interessate, inoltre, è stato evidenziato come un rapporto più stretto con tutte le fasi legate al mondo tradizionale del SE risulterebbe vantaggioso per lo sviluppo di una pipeline solida riducendo il tempo di overhead necessario per ottenere le informazioni pregnanti utili al modello.

Un primo passo importante verso delle chiare linee guida in campo di *SE4AI* è stato fatto solo recentemente, da G. Fujii et al. [23], che tramite il consorzio **QA4AI (Quality Assurance for AI-Based Products and Services)**, formato da esperti del settore e letteratura, hanno definito delle prime linee guida volte a favorire lo sviluppo di sistemi AI sempre più qualitativi.

Da questo consorzio sono stati definiti **5** aspetti fondamentali per la valutazione qualitativa di un sistema ML:

- *Data Integrity*, riguarda tutti gli aspetti legati ai dati, quindi bias, privacy, utilità, modo in cui sono stati generati...
- *Model Robustness*, riguarda tutti gli aspetti che valutano il comportamento di un modello come metriche, performance, cross validation, eterogeneità nei dati...
- *Sistem Quality*, riguarda la qualità dell'intero sistema come performance di sistema, criticità e frequenza di errori, manutenibilità...
- *Process Agility*, riguarda la qualità del processo di sviluppo come feedback, scalabilità, teamwork...
- *Customer Expectation*, riguarda la qualità del sistema dal punto di vista degli stakeholders come compliance, rispetto degli obiettivi, aspettative riposte...

Sulla base di questi aspetti, nasce un primo modello di sviluppo di sistemi di ML chiamato **Intelligent Experimental Integration (IXI) Model**, visibile in Figura 2.3, che cerca di identificare e riassumere i cinque aspetti fondamentali formando un vero e proprio processo di sviluppo per un sistema ML di qualità.

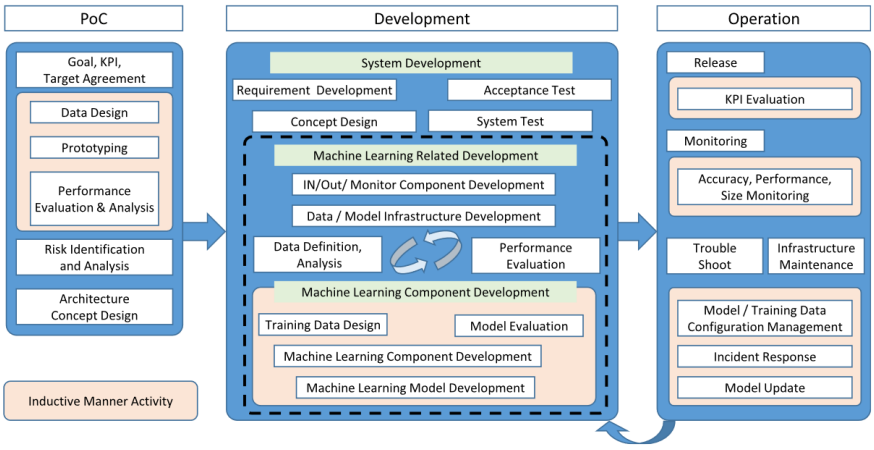

Figura 2.3: Il modello IXI per lo sviluppo di sistemi ML.

Infine, è importante citare J. Bhalla et al. [19], che in un articolo pubblicato nel 2023, cercano di definire le fondamenta di un framework, per ora solo da un punto di vista concettuale, per la *SE4AI*.

2.3 Machine Learning Fairness

Al giorno d'oggi si parla sempre di più di Intelligenza Artificiale e di Machine Learning ma questa rapida espansione è possibile solo grazie alla quantità di dati presenti nel mondo moderno. Ogni azienda genera una quantità immane di dati, anche dalle semplici operazioni quotidiane. Il Machine Learning si sviluppa verticalmente nel contesto moderno proprio grazie alla grande disponibilità e utilizzo di questi dati per poter realizzare delle previsioni e predizioni su valori futuri ma questa caratteristica pregnante crea un grande problema che tutti gli sviluppatori e gli ingegneri di ML devono affrontare, il Bias. Una definizione generale descrive il Bias come un un errore sistematico per cui il modello di ML è portato a compiere delle predizioni errate.
Il Bias esiste in diverse forme e può essere legato sia al dataset utilizzato che alle scelte effettuate in fase di progettazione ed implementazione di un modello di ML [4]. N. Mehrabi et al. [4] forniscono una classificazione del bias sulla base delle parti coinvolte, illustrato in esempio in **Figura 2.4**.

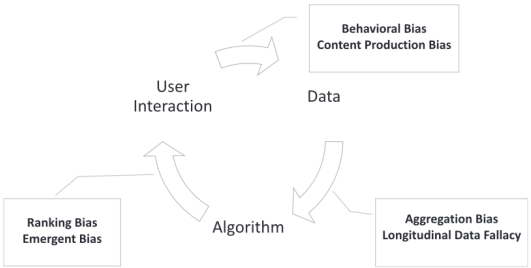

Figura 2.4: Esempi di bias nelle varie interazioni fra algoritmo, utente e dati.

La discriminazione è un altro problema cruciale nel campo del Machine Learning e in particolare della Fairness.
N. Mehrabi et al. definiscono la discriminazione come fonte di ingiustizia che nasce dal pregiudizio e dagli stereotipi umani su alcuni attributi definiti come "sensibili" o "protetti", che può avvenire in maniera intenzionale e non intenzionale [4], mentre I. Zliobaite esplica il concetto facendo riferimento ad una definizione legislativa della discriminazione che la descrive come "un comportamento svantaggioso nei confronti di individui che appartengono ad un determinato gruppo piuttosto che alle caratteristiche individuali" [25].

N. Mehrabi et al. [4] forniscono una prima classificazione di discriminazione, *Discriminazione giustificabile*, che viene esplicitamente "giustificata" in quanto è una discriminazione indissolubile legata alle informazioni pregnanti di ogni istanza di input e viene spesso ritenuta legale; *Discriminazione non giustificabile*, un tipo di discriminazione illegale che ha bisogno di essere mitigata per realizzare un modello fair. In generale, però, nei riferimenti [4, 25, 26], viene riconosciuta principalmente una divisione del concetto di discriminazione in due categorie distinte, *discriminazione diretta* e *discriminazione indiretta*.

- *Discriminazione diretta*, si verifica quando un individuo riceve un trattamento svantaggioso per via di un attributo protetto [26], oppure, quando un attributo protetto di un individuo risulta esplicitamente in un risultato svantaggioso per quest'ultimo [4]. Tipicamente questo tipo di discriminazione ricade su degli attributi che vengono già identificati dalla legge come attributi su cui è illegale discriminare. Questi attributi sono spesso esplicitati e protetti dalla legislazione, ad esempio in Europa, nell'art. 21 dell'**EU Charter of Fundamental Rights**[1].

- *Discriminazione indiretta*, questo tipo di discriminazione è meno evidente, e a prima vista può risultare invisibile. In questo tipo di discriminazione, ogni individuo sembra ricevere un trattamento "equo" sulla base dei soli attributi non protetti, ma, in realtà, gli individui di gruppi protetti possono ricevere un trattamento ingiusto come effetto implicito dei loro attributi protetti [25].

Queste classificazioni ci permettono di giungere alla conclusione che la discriminazione non assumente significato prettamente negativo, in quanto, in situazioni limite, è possibile che il comportamento assunto dal modello nel trattare le istanze sia in realtà "corretto" nel mondo in cui opera ma che in generale però è necessario attuare strategie volte a rimuovere tutti quelle "vere" discriminazioni, dirette o meno, che possono presentarsi.

[1] `http://fra.europa.eu/en/eu-charter/article/21-non-discrimination`

Definiti i concetti di Bias e Discriminazione è doveroso introdurre il concetto di Fairness poiché questo concetto si lega indissolubilmente a quest'ultimi, infatti, possiamo ritenere come "fair" qualsiasi comportamento che non presenti forme di discriminazione o bias nella progettazione e realizzazione di un modello e nell'output prodotto da quest'ultimo in presenza di nuove istanze di input.
Sono tanti gli esempi di algoritmi e modelli che hanno presentato problemi di Fairness sin dal primo sviluppo ed utilizzo. L'esempio più importante, citato in qualsiasi studio di fairness, è il problema di discriminazione del sistema COMPAS [2], sistema utilizzato negli Stati Uniti che fornisce delle stime e predizioni sulla recidività dei soggetti. Questo sistema presenta sistematicamente errori nelle predizioni per via dei bias presenti in attributi sensibili come sesso e razza che porta il sistema stesso a creare discriminazione tramite stereotipi di genere e razziali [27]. Dare una definizione univoca di Fairness è impossibile, poiché quest'ultima può presentarsi nelle diverse fasi di progettazione e sviluppo di un modello di ML.
Possiamo parlare di fairness di tipo statistico quando, utilizzando le metriche di valutazione generate a partire dalla matrice di confusione, possiamo trarre delle conclusioni sul comportamento del modello [28].
Possiamo parlare di fairness anche tramite un ulteriore strumento, i grafi causali, grafi aciclici, direzionati, in cui i nodi rappresentano attributi dell'istanza e gli archi rappresentano le relazioni fra gli attributi. Questi grafi sono utilizzati per costruire dei modelli fair di ML [29].

[2] https://en.wikipedia.org/wiki/COMPAS_(software)

In particolare, ad oggi, possiamo raggruppare le definizioni di fairness in tre gruppi fondamentali:

- **Individual Fairness**, fornire predizioni simili ad individui simili.
- **Group Fairness**, trattare in maniera equa i diversi gruppi.
- **Subgroup Fairness**, quest'ultimo gruppo unisce nozioni di Individual Fairness e Group Fairness per ottenere un risultato migliore. Spesso seleziona una definizione di Group Fairness per testare se tale definizione possa valere anche in una collezione più grande.

Questa classificazione è infatti utilizzata in riferimenti come [4, 28] per classificare le diverse definizioni di fairness presentate, utili per approfondire ulteriormente le varie definizioni di fairness ad oggi individuate.
Come viene evidenziato dalla letteratura, gli argomenti di Fairness e Quality di un modello di ML sono ormai indissolubilmente legati, in un articolo del 2020, S. Biswas et al. [30], vengono utilizzate diverse implementazioni di modelli di ML per valutare la qualità di quest'ultimi prima e dopo aver svolto operazioni di fairness volte a mitigare i possibili problemi di equità presenti nel dataset e nelle decisioni del modello. Questo articolo sposta perfettamente i temi principali dello studio è stato ritenuto un ottimo punto di partenza per poter arricchire ulteriormente l'argomento.

2.3.1 Strategie per costruire un modello fair di ML

Sono state definite nel tempo diverse tecniche mirate a costruire dei modelli fair di ML. Una prima strategia per costruire un modello fair è quello di identificare e correggere eventuali bias presenti all'interno del dataset o delle decisioni.
Bellamy et Al. [31] descrivono una vera e propria fair pipeline, riportata visivamente in **Figura 2.5**, utilizzata nell'articolo per descrivere le basi di partenza di un processo di produzione di un modello fair di ML.

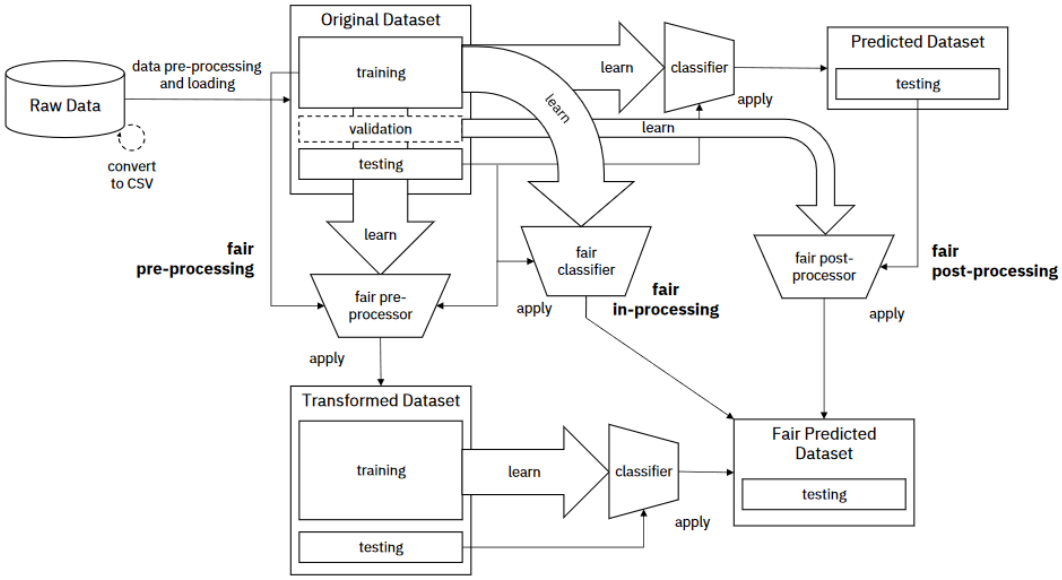

Figura 2.5: La fair pipeline definita da Bellamy et al. utilizzata per realizzare un modello fair di Machine Learning.

Dalla pipeline precedente è possibile identificare tre momenti fondamentali nel processo, in cui N. Mehrabi et al. [4] classificano le diverse strategie:

- **Pre-processing**, queste tecniche cercano di trasformare i dati di input in modo da rimuovere qualsiasi tipo di discriminazione presente.
- **In-processing**, tecniche che cercano di modificare il comportamento base degli algoritmi di apprendimento in modo da rimuovere possibili discriminazioni durante il processo di apprendimento. Queste strategie sono possibili solamente se è possibile modificare il comportamento degli algoritmi di apprimendimento,

in cui di solito, vengono incorporate delle possibili modifiche nella funzione obiettivo oppure imponendo delle restrizioni.

- **Post-processing**, queste strategie sono utilizzate quando il modello realizzato è trattato come una black box in cui non è possibile accedere e modificare i dati di input o gli algoritmi di apprendimento, in questo caso si sfruttano queste strategie che, sulla base delle etichette fornite dal modello in output, vengono riorganizzate sulla base di ulteriori algoritmi di post-processing volti ad eliminare eventuali discriminazioni generate dalle due fasi precedenti.

Questa classificazione è importante ed è stata utilizzata in più studi come [30, 31], in cui viene esplicitamente utilizzata per classificare le strategie fairness proposte e utilizzate.

2.4 Sustainability

Con l'evoluzione costante dei sistemi e artefatti software, che viaggiano verso prodotti sempre più di qualità, negli ultimi anni anche il concetto di sostenibilità del software è diventato fondamentale nel campo del Software Engineering.
Parliamo di **Sustainable Software Engineering (SSE)** come il processo di creazione di software che possa soddisfare i bisogni del presente senza compromettere la possibilità delle future generazioni di soddisfare a loro volta i propri bisogni [32].
Per sostenibilità, spesso ingenuamente ritenuta come unica definizione di sostenibilità, intendiamo l'impatto che l'artefatto software prodotto ha sull'ambiente in termini di risorse energetiche e sprechi ambientali, ma definire il concetto di sostenibilità non è semplice in quanto, nei riferimenti [32, 33], si evince che il concetto stesso di sostenibilità sia multidimensionale, diviso nei cosiddetti "pilastri", uno per ogni possibile aspetto di sostenibilità - economica, sociale, ambientale, tecnica ed individuale.
In particolare, dallo studio effettuato da S. McGuire et al. [32], si evidenzia come la sostenibilità sia un concetto multidimensionale stratificato, che permette di identificare in ogni pilastro vari livelli di specializzazione, in cui il concetto stesso di sostenibilità assume diversi significati. Un artefatto, quindi, può essere definito come sostenibile in un determinato livello ma insostenibile in altri.

Per sostenibilità economica intendiamo la capacità dell'artefatto software di rimanere economicamente efficiente nel tempo; per sostenibilità sociale intendiamo l'impatto che l'artefatto software ha sulla comunità, mentre per sostenibilità individuale, l'impatto che l'artefatto ha sul singolo individuo. Infine per sostenibilità tecnica intendiamo la robustezza, la manutenibilità e la capacità di evoluzione dell'artefatto nel tempo.
Un artefatto è, quindi, sostenibile quando ha un impatto minimo, o massimo, in ognuna di queste dimensioni e questi pilastri sono diventati degli indicatori fondamentali nel valutare l'impatto di un artefatto software dal punto di vista della sostenibilità [33].
Dallo studio condotto da S. McGuire et al. [32], vengono evidenziate delle caratteristiche negli studi condotti sulla sostenibilità, in particolare:

- il trend degli studi è di concentrarsi, fondamentalmente, sullo studio dell'efficienza energetica dell'artefatto software senza discutere in maniera olistica il concetto.
- la maggior parte degli studi tende a concentrarsi sulla sostenibilità del *prodotto* rispetto al *processo di produzione* di quest'ultimo.
- la sostenibilità è stata definita come difficile da misurare e quantificare.

Possiamo assumere quindi che, al giorno d'oggi, gli studi rilevanti in materia si concentrino principalmente su uno dei cinque pilastri.
Nel campo dell'AI e, in particolare del Machine Learning, l'interesse per la sostenibilità è di recente sviluppo e, ad oggi, gli studi più importanti riguardano il consumo energetico e la produzione di CO2 in modelli di Machine Learning e Deep Learning [34, 35].

Nello studio condotto da C. Nicodeme [36] viene evidenziato il grande dispendio di energia richiesta dallo sviluppo costante di nuovi sistemi AI.
In particolare, si fa riferimento ai consumi energetici che vengono sostenuti per conservare la grande quantità di dati richiesti dal processo iterativo di apprendimento di un modello, che prevede l'utilizzo di quanti più dati possibili per poter descrivere al meglio tutte le possibili situazioni e prevedere quanto più possibile gli *outlier*.

Possiamo quindi concludere che, in generale, gli studi più importanti sulla sostenibilità si concentrano maggiormente su uno dei 5 pilastri descritti, quello ambientale, proponendo diverse possibili soluzioni teoriche atte a ridurre i consumi energetici dei modelli e sistemi AI.

Sulla base di questa affermazione nasce uno degli obiettivi fondamentali di questo studio, cercare come sia possibile sfruttare le conoscenze attuali per fornire valutazioni concrete e tangibili di un modello negli altri ambiti della sostenibilità.

CAPITOLO 3

Metodologia

In questo capitolo verrà presentato l'obiettivo principale dello studio, le motivazioni di quest'ultimo, verranno presentate le Research Questions poste e verrà descritto formalmente il processo e gli strumenti utilizzati per rispondere ai suddetti quesiti.

3.1 Motivazioni e obiettivi dello studio

Data la grande evoluzione e il costante interesse del mondo sui sistemi AI dovuti, in particolare, ai sistemi AI rivoluzionari come GPT-3 o tool di generazione immagini come DALL-E, spesso, molti degli aspetti "secondari" dello sviluppo di un modello di ML o DL vengono completamente offuscati dai risultati ottenuti dal sistema.
L'AI, però, non è solo risultati eclatanti e sistemi che rivoluzionano il mondo, infatti, sistemi di Machine Learning, da sempre, sono utilizzati in contesti aziendali per fornire supporto ai processi. In questo campo, questi aspetti "secondari" spesso dimenticati sono fondamentali, in quanto, ci sono costi da sostenere per lo sviluppo di un modello di ML e non tutti i modelli realizzati possono realmente adempiere al compito previsto o, addirittura, fornire un reale miglioramento al processo già esistente. Questi aspetti aprono un mondo completamente differente rispetto ai risul-

tati ottenuti dai sistemi AI più famosi e, in generale, negli ultimi anni si è iniziato a parlare e valutare maggiormente questi aspetti, il più importante su tutti, il concetto di sostenibilità del prodotto, che se da un lato troviamo risultati eclatanti e sistemi che rivoluzionano il mondo, dall'altro incontriamo delle spese energetiche importanti che, in un mondo particolarmente in difficoltà con le risorse disponibili, rappresenta una vera e propria problematica per l'evoluzione dell'AI e per la sostenibilità del nostro pianeta.
Nonostante il grande interesse in letteratura per queste tematiche, al giorno d'oggi, non sono presenti degli studi importanti che evidenzino i trade-off qualitativi effettuati quando si applicano tutte le strategie di equità già ben definite e, in particolare nell'ambito della sostenibilità, l'assenza di studi che valutino le implicazioni sociali, economiche ed ambientali della produzione di un modello "equo".

Questo studio nasce proprio dall'interesse sempre più crescente verso le tematiche di "equità" e "sostenibilità" dei sistemi AI e, nello specifico, di modelli di Machine Learning e vuole essere un punto di inizio per delle valutazioni sulla sostenibilità, non solo ambientale ed energetica, dello sviluppo di un sistema di ML "equo".

Prima di presentare i quesiti posti è quindi importante formulare in maniera rigorosa l'obiettivo principale di questo studio.

◎ **Our Goal.** Valutare come le tecniche di Fairness influenzino la Qualità e la Sostenibilità dei sistemi di Machine Learning

Fissato l'obiettivo principale, di seguito verranno elencate le **Research Questions** individuate con le proprie motivazioni:

🔍 **RQ$_1$.** *Che impatto hanno gli strumenti di Fairness sulla sostenibilità sociale del prodotto?*

In questa prima RQ cerchiamo di definire se i modelli "fair" realizzati possano portare benefici a livello sociale aziendale ed individuale, in particolare, tramite i risultati ottenuti, si vogliono valutare i modelli ottenuti sia dal punto di vista qualitativo che etico, fornendo delle possibili implicazioni sui benefici sociali che un prodotto

può portare ad un individuo soggetto alla predizione del modello e all'azienda che utilizza tale modello.
In questa RQ, si vuole rispondere alla domanda *"Che benefici porta utilizzare un modello oggettivamente equo?"*, dal punto di vista individuale, valutare quindi se la decisione di un modello "equo" possa influire, positivamente o negativamente, sull'individuo soggetto della predizione del modello; dal punto di vista aziendale, valutare l'apporto, positivo o negativo, che un modello possa portare al processo e al problema per cui è stato realizzato.

RQ$_2$. *Che impatto hanno gli strumenti di Fairness sulla sostenibilità ambientale del prodotto?*

In questa RQ si vuole valutare il consumo energetico richiesto per produrre diversi modelli di ML per poter valutare e confrontare i risultati ottenuti e scegliere un soluzione "ottima" al problema posto in essere.
In particolare, si vuole valutare il "peso" di tutte le operazioni, volte a favorire lo sviluppo di un modello qualitativamente migliore ma anche oggettivamente "equo", sfruttando tutte le classiche operazioni di training e testing di un modello standard e di pre e post processing offerti dai tools moderni di fairness volte a rimuovere la discriminazione e favorire un risultato "equo".

RQ$_3$. *Che impatto hanno gli strumenti di Fairness sulla sostenibilità economica del prodotto?*

In questa RQ si vuole racchiudere un po' una somma di tutte le osservazioni precedenti per definire il possibile impatto economico che realizzare un modello fair ha in termini di efficienza del modello, praticità, richiesta di risorse, predizioni oggettivamente corrette.
Questa RQ è molto importante poiché permette di avere una visione generale di sostenibilità del prodotto; ci permette di individuare quali sono i trade-offs qualitativi necessari che puntano a rimuovere quanta più discriminazione dai dati sensibili e mantenere una qualità predittiva elevata.

Per formulare una risposta rigorosa a queste RQ è necessario analizzare e descrivere formalmente il processo e gli strumenti utilizzati attraverso le *variabili dipendenti* e le *variabili indipendenti* del processo.

3.2 Variabili indipendenti dello studio

3.2.1 I datasets

Per poter realizzare i modelli sono stati individuati **5** datasets in cui sono presenti attributi di tipo sensibile.
La scelta di utilizzare più datasets permette di realizzare molti più modelli esplorando anche diversi settings di dati. In particolare, il filo conduttore della scelta dei datasets è stata la presenza di attributi ritenuti sensibili su cui poter testare eventuale presenza di discriminazione, sia intrinseca dei dati che generata dai modelli in fase di predizione. Un'alta caratteristica fondamentale di alcuni di questi datasets, in particolare l'Adult Dataset e il German Credit Dataset, è la presenza di numerosi studi sulla Fairness che utilizzano quest'ultimi, è una caratteristica fondamentale che permette di effettuare anche delle correlazioni e incrociare i risultati ottenuti, nella descrizione in seguito verranno citati alcuni lavori di Fairness in cui tali datasets sono stati utilizzati.
La scelta di utilizzare più datasets ricade nel cercare di analizzare dei modelli ottenuti da diversi settings di dati, in particolare, sono stasti scelti datasets quanto più diversificati sia nei temi trattati che nel numero di features presenti e nella grandezza della popolazione.
Questo permette di ottenere risultati da modelli sostanzialmente diversi, che adempiono a compiti diversi, utilizzando un diverso quantitativo di informazioni disponibili.
Vengono presentati ora i 5 datasets con alcune delle loro caratteristiche:

- **Adult Dataset** [37], un grande dataset sfruttato costantemente in ambito Fairness. In questo dataset si vuole predire l'appartenenza ad una specifica classe di reddito annuale di un individuo, in particolare vengono classificati nel dataset individui il cui reddito supera i 50 mila euro annuali e individui il cui reddito è

inferiore ai 50 mila euro annuali. In questo dataset è importante il numero di entry presenti, parliamo di quasi *50.000* diverse entry, contando anche le entry che presentano valori nulli, che formano una popolazione importante ai fini di predizione, descritte da **14** features che contengono valori come informazioni sul settore di impiego, tipo di impiego, ore di lavoro settimanali...
In questo dataset non mancano attributi sensibili, fra cui *Età*, *Sesso*, *Razza* e attributi personali sui quali il risultato potrebbe presentare discriminazione e favoritismi come *Educazione* e *Stato Civile*. È quindi una scelta molto valida per lo studio che permette di fornire una grande varietà di risultati e di possibili outcomes. Alcuni studi di Fairness in cui questo dataset è stato utilizzato: [38, 39, 40]

- **German Credit Dataset** [41], dataset il cui scopo è predire la possibilità che un cliente bancario ha di ripagare, o meno, un prestito. In questo dataset sono presenti *1000* istanze di clienti, descritte da **20** diverse features, che spaziano in valori categorici e numerici e riguardano informazioni come . In particolare, sono presenti *2* possibili attributi protetti, *Età* e *Sesso*. La scelta di questo dataset ricade sull'enorme utilizzo di quest'ultimo nel campo della Fairness, è quindi più che giusto sfruttare un dataset già ampiamente utilizzato in letteratura per poter ottenere dei risultati veritieri e trovare corrispondenza negli studi associati. Alcuni studi di Fairness in cui è stato utilizzato questo dataset: [42, 43, 44].

- **Home Credit Default Risk Dataset** [45], dataset contenente informazioni reali legate ad informazioni bancarie di diversi individui collezionati da diverse banche. Questo dataset è stato reso disponibile per una delle tante Challenge presenti sulla piattaforma **Kaggle**, una piattaforma cardine del Machine Learning, Deep Learning e Data Analysis. Questo dataset contiene il maggior numero di entry, sono presenti circa *100* mila entrate differenti per un totale di oltre **120** features per ogni entrata.

L'obiettivo di questo dataset, è prevedere se un cliente sia in grado di sanare un eventuale prestito bancario di grande importanza, in questo caso per l'acquisto di un immobile ad uso residenziale. Non mancano features ritenute sensibili e personali, come *Età*, *Sesso*, *Stipendio Annuale* e cosi via.
Il grande numero di features ed di entrate, nonchè la realtà dei dati raccolti, sono caratteristiche fondamentali di questo dataset permettono di effettuare operazioni e valutazioni di fairness non indifferenti su dati reali, valutando il comportamento degli strumenti di fairness e dei modelli anche su una possibile applicazione reale.

- **Heart Disease Dataset** [46], dataset che presenta informazioni di tipo medico legate a diversi pazienti nel campo delle malattie al cuore. L'obiettivo del dataset è proprio predire, date le informazioni rilevati per ogni paziente, se è possibile valutare la presenza o meno di possibili infezioni o malattie al cuore. Principalmente il dataset fornisce informazioni sull'assenza o meno e sul grado di severità possibile della malattia. Queste informazioni, in letteratura, sono spesso utilizzate per considerare la presenza o meno di una malattia o infezione, tralasciando il grado di severità. Contiene circa *300* istanze, caratterizzate da **13** features fondamentali ai fini della classificazione. Essendo un dataset medico non mancano attributi sensibili legati al *Sesso*, *Età*, *Malattie pregresse* e tutte quelle informazioni legate alla storia clinica del paziente.
  Questo dataset ci permette di espandare gli studi e le valutazioni di Fairness anche in campo medico, campo in cui questo tipo di applicazione rappresenta sicuramente una soluzione critica ed è quindi necessario garantire un certo livello di qualità, sicurezza e validità delle soluzioni proposte.

- **UTKFace** [1], ultimo dataset, l'obiettivo è predire il sesso dell'individuo. Si discosta dai datasets descritti in precedenza, in quanto questo dataset contiene immagini, piuttosto che informazioni, relative ad individui di sesso maschile e femminile, che variano fra 4 possibili razze e spaziano nel range da 0-116 anni di età.

[1] `https://susanqq.github.io/UTKFace/`

Nel dataset è presente una popolazione formata da circa *25.000* entrate.
Il dataset presenta, come i precedenti, attributi sensibili fra cui *Età* e *Razza*.
Questo dataset è stato scelto per espandere superficialmente lo studio di Fairness anche a modelli realizzati tramite reti neurali di DL e non solo a modelli di ML.

L'Adult Dataset e il German Credit Dataset sono reperibili sulla piattaforma *UCI Machine Learning*[2], mentre l'Home Credit Default Risk è presente sulla piattaforma *Kaggle*[3], infine l'UTKFace Dataset è disponibile alla propria pagina.[4]

3.2.2 I gruppi protetti e non protetti

Per dividere i dataset in gruppi protetti, o sfavoriti, e gruppi non protetti, o favoriti, è stato scelto di utilizzare almeno due attributi comuni per tutti i dataset. La scelta ricade sugli attributi **Sesso** ed **Età**, presenti in tutti i dataset, ad eccezione dell'UTKFace Dataset in cui uno di questi due attributi è proprio il target, attributi che spesso sono stati utilizzati in letteratura negli studi di Fairness più famosi, ottimi per valutare realmente discriminazioni all'interno del dataset e delle predizioni dei modelli.
Sono stato scelti come gruppi non protetti tutte le entrate di sesso maschile o con un età uguale o superiore alla media del dataset. Il gruppo protetto, invece, contiene tutte le entrate di sesso femminile e con età minore della media del dataset. Queste particolari combinazioni di valori sugli attributi sensibili sono state scelte per poter valutare eventuali discriminazioni di genere o di età, in campi come l'economia e la salute, ampiamente trattati dai datasets scelti.
In fase di mitigazione verranno considerati i due sottogruppi come un unico gruppo protetto da fornire agli strumenti di Fairness individuati, andando poi, in fase di valutazione dei risultati, a distinguere i valori ottenuti per i singoli sottogruppi.

[2] `https://archive.ics.uci.edu/datasets`
[3] `https://www.kaggle.com/competitions/home-credit-default-risk/overview`
[4] `https://susanqq.github.io/UTKFace/`

Per l'UTKFace Dataset, invece, è stato utilizzato l'attributo **Razza**, un attributo chiave per un modello di riconoscimento di immagini, in quanto, spesso, il diverso colore della pelle può condizionare le scelte del modello nel produrre una predizione. Infine, nelle **Tabelle 3.1,3.2** vengono presentate delle caratteristiche pregnanti dei datasets, come popolazione effettiva, in quanto sono state escluse le entries che presentavano valori nulli, numero di features realmente utilizzate per la predizione, numero di predizioni positive e negative per i gruppi favoriti e non, che possono fornire un primo quadro generale della situazione presente nei dataset prima delle possibili valutazioni di fairness. Nelle tabelle vengono utilizzati gli acronimi **GNP** e **GP** per indicare "Gruppi Non Protetti" e "Gruppi Protetti", inoltre, in **Tabella 3.2**, viene indicato con **Y** il valore della variabile *Target*, ovvero la variabile da predire, e con **Y=1** e **Y=0** le predizioni positive e negative rispettivamente, questo sia per i "Gruppi protetti" (GP) che per i "Gruppi Non Protetti" (GNP).

Dataset	Popolazione	Features	Entries GP	Entries GNP
Adult Dataset	30162	14	5670	24492
German Credit Dataset	1000	20	319	681
Home Credit Dataset	99300	60	32637	66663
Heart Disease Dataset	297	14	39	258
UTKFace Dataset	24104	2	6269	17835

Tabella 3.1: I dataset scelti e le loro caratteristiche

Dataset	Y=1 GP	Y=0 GP	Y=1 GNP	Y=0 GNP
German Credit Dataset	195	124	505	176
Adult Dataset	439	5231	7069	17423
Home Credit Dataset	16980	15657	32670	33993
Heart Disease Dataset	2	37	135	123
UTKFace Dataset	3164	3105	8358	9477

Tabella 3.2: I dataset scelti e alcune caratteristiche sui gruppi protetti e non protetti

3.2.3 Gli algoritmi di ML e le reti neurali di DL scelte

Per poter analizzare quanto più possibile i diversi datasets e i possibili outcomes è stato scelto di utilizzare 4 algoritmi di ML e 2 reti neurali di DL per poter confrontare i risultati ottenuti e avere una visione completa del contesto di ogni dataset.
In particolare, per i primi 4 dataset sono stati utilizzati ben **4** algoritmi diversi di ML. **LogisticRegression**, un algoritmo che sfrutta il concetto di regressione; **RandomForestClassifier**, algoritmo che sfrutta la potenza degli alberi per generare una predizione; **LinearSVC**, algoritmo che sfrutta le *Support Vector Machine (SVM)* ed infine **XGBoostClassifier**, che sfrutta la tecnica del *gradient boosting* sugli alberi.
Le implementazioni dei primi 3 algoritmi sono presenti nella libreria di **Scikit-Learn**, disponibile per più linguaggi di programmazione mentre il modello di XGBoost è fornito dalla libreria omonima **XGBoost**, che implementa diverse soluzioni sfruttando l'omonimo algoritmo.
Per l'ultimo dataset, sono state utilizzate due reti neurali particolarmente usate in campo di image recognition, il **MobileNetV2**[5] e **ResNet50**[6], utilizzando le versioni delle reti implementate nella libreria **TensorFlow**, disponibili anche nella piattaforma **Kaggle**.

[5] `https://www.kaggle.com/models/tensorflow/mobilenet-v2`
[6] `https://www.kaggle.com/models/tensorflow/resnet-50`

3.2.4 I tool di Fairness utilizzati

Per poter valutare Fairness e operare sui datasets e modelli sono state utilizzate le due librerie di fairness presenti al momento sul mercato, la libreria **AIFairness360** [31] e la libreria **FairLearn** [47].

AIFairness360 nasce come framework per poter valutare datasets e modelli su diverse definizioni di Fairness e fornire strumenti utili a rimuovere discriminazione in pre, in e post-processing.

È un framework in costante e continua crescita, di proprietà della IBM, una delle pietre miliari dell'informatica. Offre più di 70 metriche di valutazione sulla base delle definizioni più comuni di Fairness presentate in letteratura, e diversi strumenti per la mitigazione di diversi tipi di bias e discriminazioni.

Nello studio, sono state utilizzate soluzioni offerte da questa libreria in tutti i campi di pre, in e post-processing, utilizzando le implementazioni base fornite, senza modificare parametri opzionali, in modo da evidenziare fortemente la replicabilità e la semplicità di utilizzo della libreria e i risultati prodotti.

In particolare, è stato utilizzato l'oggetto **Reweighing** per la fase di pre-processing, che permette di modificare il peso attributo alle singole istanze del datasets, individudando le istanze appartenenti a gruppi protetti e non, cercando i pesi ideali volti a mitigare qualsiasi tipo di discriminazione presente all'interno del dataset.

Questo oggetto è stato scelto come soluzione di pre-processing, in quanto permette di lavorare su gruppi protetti formati dall'unione di più sottogruppi ed operare su diversi attributi sensibili senza necessariamente trattare i singoli sottogruppi.

Per la fase di in-processing, è stato utilizzato l'oggetto **MetaFairClassifier**, una implementazione generale di un classificatore che tiene conto, oltre alle caratteristiche pregnanti delle features, anche le metriche di fairness indicate. Quest'oggetto è stato sfruttato in fase di in-processing per modificare il set di training standard in un set di training che tenga conto delle metriche di Fairness indicate in precedenza, una volta modificato il training set, vengono addestrati dei modelli standard ulteriori sul nuovo training set modificato, in maniera da fornire una fase di training che utilizzi principalmente dei training set più fair.

Come per l'oggetto precedente, questa soluzione di in-processing permette di lavorare con più sottogruppi protetti e attributi sensibili in maniera rapida e immediata, ottentendo da subito risultati senza necessariamente esaminare singolarmente i diversi attributi sensibili.
Infine per la fase di post-processing, è stato utilizzato l'oggetto **EqOddsPostprocessing**, oggetto che permette di calibrare le predizioni del modello cercando di garantire delle predizioni che tengano conto di Fairness dei gruppi individuati.
Questo oggetto risulta il più realistico fra le varie soluzioni proposte dalla libreria, in quanto modifica gradualmente i valori ottenuti in maniera da massimizzare le metriche piuttosto che applicare un singolo procedimento, come l'inversione completa delle predizioni dei vari gruppi, e valutarne la Fairness ottenuta.
Inoltre, tutte queste strategie sono state utilizzate ed implementate sia per i modelli di ML che i modelli di DL.
FairLearn, nasce come framework per valutare diverse metriche di Fairness basate sulle definizioni più comuni presenti in letteratura e fornire strumenti grafici in grado di visualizzare e mappare tali metriche in grafici e reports che presentano in maniera lampante le eventuali situazioni di discriminazione e unfairness presenti nel datasets o nei modelli.
Anche questa libreria è in costante aggiornamento e sviluppo, d'altronde, il progetto è finanziato sin dagli albori dalla Microsoft, che come l'IBM, rappresenta un'altra pietra miliare dell'informatica moderna.
Anche questa libreria, nonostante nasca come libreria incentrata su metriche e sulla rappresentazione visiva di quest'ultime, offre diverse soluzioni di pre, in e post-processing su datasets e modelli in grado di mitigare bias e discriminazioni di vario genere ed, anche in questo caso, sono state sfrutta le soluzioni proposte con le loro implementazioni base, senza modifiare parametri opzionali, per evidenziare facilità di utilizzo, replicabilità e immediatezza dei risultati. In particolare, nello studio sono stati usati gli oggetti di **CorrelationRemover**, oggetto di pre-processing che punta a togliere il tasso di correlazione presente nel dataset fra gli attributi ritenuti sensibili e i rimanenti attributi; **ThresholdOptimizer**, oggetto di post-processing in grado di valutare le scelte del modello indicato e in maniera *graduale* ripetere queste predizioni fino ad ottenere risultati equi secondo determinate metriche indicate.

Purtroppo, per via dell'implementazione dell'oggetto **ThresholdOptimizer**, che richiede un oggetto di tipo *estimator*, non è stato possibile implementare questa soluzione anche per la controparte di DL, in quanto incompatibile con i modelli della libreria **Tensorflow**. Verrà quindi utilizzato il solo oggetto **CorrelationRemover**, di tipo pre-processing, per effettuare mitigazione anche sulla parte di DL.
Entrambe le soluzioni scelte di questa libreria, ad oggi, sono le uniche offerte dalla libreria stessa nei due campi trattati.
Entrambe le librerie sono molto valide in termini di Fairness e di operazioni possibili. Sono dotate di guide, docs e esempi pratici molto dettagliati che permettono a chiunque di interfacciarsi con il mondo della Fairness senza particolari necessità.
Tutti gli oggetti presentati e scelti dalle due librerie, inoltre, opereranno ricevendo in input contemporaneamente entrambi i sottogruppi protetti evidenziati dagli attributi sensibili scelti come un un unico gruppo. Verranno, poi, considerati i risultati ottenuti singolarmente per ogni sottogruppo per fornire ulteriori informazioni sull'operato di ogni soluzione.
Nello studio, in particolare, verranno valutati i risultati ottenuti per concludere sia un confronto fra le possibili interazioni fra queste due librerie, ove possibile, sia dei risultati e modelli prodotti per poter valutare equità e sostenibilità dei modelli stessi.

3.2.5 Il tool di Sustainability utilizzato

Per valutare la sostenibilità energetica è stata utilizzato il framework **CodeCarbon**[7], che, tramite la libreria omonima offerta, permette di generare una stima sulla quantità di CO2 prodotta dall'esecuzione di parti di codice, funzioni e scripts.
È un framework molto recente, in costante sviluppo ed espansione, data la natura stessa dell'argomento sempre più d'interesse. Offre strumenti in grado di visualizzare i risultati ottenuti e compararli con i consumi moderni sia mondiali che territoriali. È, inoltre, uno strumento molto valido, in quanto permette di calcolare consumi energetici del codice senza particolari necessità di setting di ambiente e codice, ed offre la possibilità di valutare interi snippets di codice, funzioni e scripts senza particolari requisiti.
Offre la visualizzazione di tutti i risultati anche in maniera *cloud*, in grado di aggiornarsi costantemente alle ultime esecuzioni di codice per poter valutare in *real-time* tutti i consumi.
Nello specifico, offre diverse implementazioni possibili e pattern dell'oggetto fulcro delle misurazioni, **TrackerEmission**, in grado di ottenere informazioni sui consumi, tempi e risorse per poter generare un report finale.
Il report generato contiene una serie di informazioni pregnanti legate al consumo di energia richiesto da ogni singolo componente rilevato fra CPU, GPU e RAM ed un totale complessivo, informazioni sul totale emissioni di CO2 prodotta, in KG, e sul rateo di emissione, nonchè ulteriori informazioni di contorno sul sistema operativo, tempo richiesto, componenti e cosi via.
Ha permesso allo studio di valutare tutte le fasi della creazione di un modello, partendo dalla fase di analisi del dataset, manipolazione di quest'ultimo fino ad arrivare alla fase di validazione dei modelli ottenuti e generazione delle metriche qualitative e di Fairness.

[7] `https://codecarbon.io/`

3.2.6 L'ambiente di lavoro

Per realizzare l'intero studio è stato utilizzato il linguaggio **Python (vers. 3.10.12)**, sfruttando librerie di punta, come **Pandas (vers. 2.1.3)** e **Numpy (1.26.2)**, per la gestione e manipolazione di strutture come dataframes e arrays. Per l'implementazione degli algoritmi di ML, delle strategie di testing e validation dei dataset e la generazione di metriche qualitative dei modelli è stata usata la già citata libreria **Scikit-Learn (vers. 1.3.2)**. Per le reti neurali, il loro apprendimento e testing e con la corrispettiva generazione di metriche di valutazione, come già citato, è stata usata la libreria **TensorFlow (vers. 2.14.0) e Tensorflow-Addons (vers. 0.22.0)**. Per la realizzazione di grafici, plot e mappe sono state sfruttate le librerie **Seaborn (vers. 0.13.0)** e **Matplotlib (vers. 3.8.1)**. Per salvare i modelli di ML realizzati in file da poter facilmente aprire e riutilizzare i modelli è stata usata la libreria **Pickle**.
Per le due librerie di Fairness e CodeCarbon sono state utilizzate le ultime versioni disponibili ad oggi, **AIF360 (vers. 0.5.0), Fairlearn (vers. 0.9.0) e CodeCarbon (vers. 2.3.1)**. Infine, per realizzare ed eseguire tutti gli script è stato utilizzato l'editor **Visual Studio Code**.

3.2.7 La macchina di esecuzione

Tutti gli esperimenti sono stati svolti su un computer portatile dotato di processore **AMD Ryzen 7 5800H**, processore con un alto clock-rate, partendo da 3.2GHz base fino ad un massimo di 4.4GHz e **16GB, DDR4, 3200MHz** di RAM, componenti molto valide per velocizzare ed eseguire senza problemi tutte le fasi più pesanti di ML e DL, dotato, infine, di scheda video **Nvidia RTX 3060**, scheda video che, ad oggi, risulta molto valida in termini di prestazioni permettendo di addestrare modelli di DL con degli strumenti adatti. Per poter ottenere delle misurazioni di consumi, risorse e tempo impiegato, quanto più fedeli possibili alla sola esecuzione del codice tutte le esecuzioni di codice sono state effettuate in maniera isolata dalla rete, in modalità aereo e cercando minimizzare il numero di processi attivi durante le misurazioni.

3.3 Variabili dipendenti dello studio

3.3.1 Le metriche di equità

Per poter valutare la Fairness di un modello o di un dataset sono state utilizzate in particolare le metriche di **Mean difference**, **Disparate Impact** e **Equalized Odds**. Queste metriche sono standard del testing di Fairness, in quanto permettono di fornire una visione generale delle possibili discriminazioni ed situazioni di non equità presenti nei dati sui cui il modello basa la sua intera conoscenza, e quindi, il suo funzionamento.
La **Mean Difference**, conosciuta anche come **Statistical Parity Difference**, è una metrica prevalentemente legata al pre-processing dei dati, ci permette di conoscere meglio le probabilità di predizione di un modello per i gruppi "protetti" (o "non privilegiati") e i gruppi "non protetti" (o "privilegiati").
In particolare, questa metrica è cosi calcolata:

$$Mean\ Difference = Pr(Y = 1|D = unprivileged) - Pr(Y = 1|D = privileged)$$

Questa metrica, in fase di pre-processing, ci permette di individuare una possibile disparità nel numero di istanze positive per le due classi, che se presenti, possono risultare in un modello che presenta "favoritismi" nei confronti di una classe, sulla base degli attributi da proteggere ritenuti sensibili. In fase di in-processing, ci permette di individuare se il modello soffre di possibili "favoritismi" in caso di predizioni positive per le classi privilegiate o meno e di tarare in maniera "fair" i risultati ottenuti.
Questa metrica da sola non ci può fornire un quadro generale delle prestazioni del modello e delle caratteristiche intrinseche dei dati stessi, per cui è quindi necessario utilizzare un ulteriore metrica per poter fornire delle stime di equità maggiori ed analizzare anche eventuali casi di discordanza fra metriche diverse, generate da caratteristiche pregnanti dei dati stessi. Il **Disparate Impact (DI)**, è una metrica molto importante e spesso utilizzata, in combinazione ad altre metriche, in quasi tutti gli altri studi di Fairness.
Questa metrica, cosi calcolata:

$$Disparate\ Impact = \frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)}$$

È definita come il rapporto fra la porzione di istanze del gruppo non privilegiato che ricevono predizioni positive ed istanze del gruppo privilegiato che ricevono predizioni positive.
Questa metrica è rappresentabile come percentuale o come valore discreto nel range [0,1], nelle definizioni standard, si segue una regola definita come *quattro-quinti*, regola per cui se il gruppo non privilegiato riceve un risultato positivo meno dell'80% rispetto al gruppo privilegiato, allora parliamo di una violazione del Disparate Impact.
Questa metrica è valutabile sia in fase di pre-processing, valutando il numero di positivi per ognuna delle due sottoclassi e il rapporto che esiste fra essi, che in fase di in e post-processing valutando il comportamento del modello.
Infine, l'**Equalized Odds**, conosciuta anche come *Sensibilità* è una metrica di fairness che valuta l'equità di un modello predittivo rispetto a gruppi demografici distinti. Misura la coerenza delle prestazioni del modello tra questi gruppi, assicurandosi che le stesse opportunità di predizione accurata siano estese a tutti, indipendentemente dalle caratteristiche demografiche. L'obiettivo è promuovere una decisione imparziale e giusta in tutti i contesti applicativi del modello.
Questa metrica è descritta come:

$$Eq.\ Odds = P[h(X) = 1|A = a, Y = y] = P[h(X) = 1|Y = y]\ \forall a, y$$

Questa metrica è principalmente individuabile e mitigabile in fase di in e post-processing, in quanto richiede predizioni da parte del modello che vengono confrontate con i reali valori del test per poter fornire informazioni in merito alla metrica.

Queste metriche sono state scelte quindi, come è possibile evincere dalla loro descrizione, sia dall'utilizzo di indicatori simili, in quanto usano la probabilità di predizioni positive nelle due sottoclassi, che nella loro contrapposizione, in quanto il *Mean Difference* è spesso individuabile e mitigabile in fase di pre-processing, mentre, il *Disparate Impact* è mitigabile in tutte le fasi di pre, in e post-processing, mentre l'*Equalized Odds* viene maggiormente studiato e mitigato in fase di in e post-processing, queste caratteristiche rendono le metriche strettamente correlate e un buon punto di inizio per una discussione e valutazione veritiera e dettagliata della Fairness di modelli e dati.

3.3.2 Le metriche di sostenibilità

Per poter valutare i consumi di un modello e, quindi, la sua sostenibilità energetica sono state utilizzate le entries del file di output della libreria CodeCarbon relative all'esecuzione del codice indicato.
Questo file di output contiene informazioni omogenee sui consumi sostenuti dall'esecuzione di codice, spaziando dal consumo energetico richiesto dalla CPU, dalla RAM a quello richiesto dalla GPU. Lo script di creazione di questo file di report, realizza anche una stima molto importante in termini di consumo e produzione di CO2. Questo valore viene spesso definito come *Carbon Footprint*, ovvero, l'impronta lasciata a livello ambientale da un processo di produzione, in questo caso, del codice appena eseguito.
Al giorno d'oggi il concetto di *Carbon Footprint* è un tema sempre più scottante dell'informatica ma anche della sostenibilità in generale, se da una parte troviamo settori in costante e verticale crescita ed espansione da una parte troviamo consumi e scarti di produzione sempre più imponenti.
Questa stima di valore mi permette di fare delle deduzioni importanti sui consumi del ML, poiché se un modello semplice di ML per un progetto di studio di un argomento genera una quantità importante di scarti, in particolare di CO2, è chiaro come modelli realizzati per sistemi più avanzati e per situazioni reali sostengano dei costi importanti.

3.3.3 Le metriche qualitative

Per poter valutare la qualità dei modelli realizzati è stato scelto di utilizzare le metriche di **Accuracy**, **F1-Score**, **Precision** e **Recall**. Queste metriche sono spesso utilizzate contemporaneamente poiché ci ci permettono di avere un quadro generale delle prestazioni di ogni modello e confrontare quest'ultimi sulla base dei risultati ottenuti.

L'**Accuracy** è formalmente descritta come il rapporto fra il numero di predizioni corrette rispetto al numero totale di predizioni.

$$Accuracy = \frac{Predizioni\ corrette}{Predizioni\ totali}$$

Nel caso di una classificazione di tipo binario, l'accuracy è anche descrivibile come il rapporto fra Positivi e Negativi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Questo tipo di metrica, utilizzata da sola, non ci permette, però, di fornire delle stime complete sulla qualità del modello realizzato, in particolare, l'accuracy pecca di significatività in caso di set non bilanciati di dati, in quanto un numero sbilanciato di positivi o di negativi tenderà a realizzare un risultato difettoso dato dallo sbilanciamento delle classi.

In particolare, questa metrica pecca in situazioni in cui il modello è particolarmente in grado di classificare correttamente uno dei due possibili outcome, considerando un modello con predizione binaria. Per questo c'è bisogno di utilizzare ulteriori metriche per fornire un risultato concreto.

La **Precision** è una metrica che viene utilizzata per indicare quale porzione delle predizioni positive è realmente positiva.

Questa metrica, nel caso di classificazione binaria, è possibile descrivere come:

$$Precision = \frac{TP}{TP + FP}$$

La metrica **Recall**, invece, ci permette di individuare quale porzione della classe realmente positiva è stata predetta correttamente come tale. Questa metrica, in una classificazione binaria, è descrivibile come:

$$Recall = \frac{TP}{TP + FN}$$

Queste due metriche, sebbene possano sembrare simili, forniscono delle informazioni molto importanti in tutte quelle situazioni la cui sola *Accuracy* non basta per descrivere correttamente il modello.
Infine, la metrica **F1-Score** viene definita formalmente come il rapporto fra **Precision** e **Recall**, definita anche come *Media armonica* fra le due metriche, è quindi intuibile come questa metrica venga fortemente correlata al calcolo delle due precedenti. Questa metrica è descrivibile come:

$$F_1\text{-}Score = 2 \cdot \frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

Questa metrica combina le informazioni ottenute dalle due metriche precedenti per fornire risposte sulla qualità delle predizioni fornite dal modello.

3.3.4 Training e Testing dei modelli

Le fasi di Training e Testing rappresentano nel campo del ML le due fasi cruciali dello sviluppo di un modello di ML.
Esistono innumerevoli strategie di training e testing, che variano dalle più semplici ed immediate alle più dettagliate. La scelta di queste strategie è quindi piuttosto complessa, molte di queste strategie possono risultare in risultati migliori, o peggiori, a seconda delle caratteristiche pregnanti dei dati di training e di testing e degli algoritmi che si sceglie di utilizzare.
La strategia più diretta prevede la suddivisione del dataset seguendo la regola del *80/20*, 80% dei dati è dedicato al training e il 20% al testing. In generale, questa strategia può darci subito un primo possibile riscontro dell'algoritmo scelto e del modello sui dati presentati e, per questo, è stata scelta come strategia per lo studio posto in essere.
Questa strategia ci permette di fornire un immagine chiara dell'affinità che il modello realizzato su un particolare algoritmo ha sul problema posto dal dataset e permette facilmente di valutare se gli algoritmi scelti sono corretti o meno.

Questa particolarità è molto importante per lo studio, poiché ci permette di fornire informazioni anche sulla correttezza della scelta dell'algoritmo e cosa comporta una scelta "corretta" di un algoritmo rispetto ad una scelta "sbagliata", in termini di sostenibilità sociale, individuale ed economica.

3.3.5 L'analisi dei risultati

Tutti i risultati prodotti, presenti nei vari reports generati dall'esecuzione di ogni, verranno utilizzati e commentati per produrre una risposta quanto più dettagliata per ogni RQ individuata. In particolare, tutte le metriche, di qualità e di Fairness, verranno confrontate statisticamente di volta in volta con i valori ottenuti dai modelli standard, che non presentano alcun tipo di operazione di Fairness. Inoltre, le metriche di Fairness verranno considerate ed analizzate per ogni sottogruppo protetto evidenziato. Verranno accostati quanto più possibile i risultati prodotti dalle due librerie con le loro soluzioni, andando ad evidenziare, tramite un confronto percentuale ed analitico, eventuali miglioramenti e peggioramenti delle metriche riportate. Questi valori verranno utilizzati, esaminando la tendenza statistica al miglioramento o peggioramento delle metriche, per valutare positivamente o negativamente la strategia offerta da ogni libreria ed, nel complesso, la strategia di pre, in o post-processing in termini di qualità ed equità.
I valori ottenuti dalle misurazioni in merito a consumi energetici, produzione di CO2 verranno valutati su una base comune di tempo, andando quindi a valutare, nell'arco di un'ora, quali sarebbero i consumi di un esecuzione di ogni script in questo lasso di tempo, cosi da fornire una visione molto dettagliata, su grandezze simili, per poter fornire una risposta concreta alla RQ2.
Inoltre, viene valutato il tempo medio di esecuzione dei vari scripts, per poter analizzare la tendenza al miglioramento o peggioramento delle tempistiche, valutando l'impatto di ogni soluzione proposta in termini di tempo necessario aggiuntivo per ottenere un prodotto maggiormente "equo". Infine, vengono valutati anche i pesi dei modelli ottenuti in MB, per analizzare quale algoritmo di ML e DL sia il migliore ed evenutali aumenti dei pesi in base alle operazioni di Fairness sostenute.

3.4 La struttura del progetto

Il progetto, disponibile sul mio profilo GitHub [8], presenta diverse directory che racchiudono datasets, scripts di esecuzione e reports.
In particolare, per ogni dataset, è presente una cartella in cui è contenuto uno script in cui vengono addestrati dei modelli sul dataset *as-is* (senza modifiche), e 2 scripts per libreria, *AIFairness 360* e *FairLearn*, distinti per il tipo di operazione (pre, in e post-processing) effettuata.
Ogni script addestra **4** modelli *standard* sfruttando i **4** algoritmi descritti in precedenza.
Per ogni libreria è presente uno script che effettuando operazioni di pre-processing sul dataset permette di ottenere un dataset modificato sul quale vengono addestrati **4** nuovi modelli sfruttando sempre gli stessi algoritmi.
Questa scelta è stata fatta per confrontare i risultati ottenuti da un modello sul dataset base e da un modello ottenuto dal dataset rimodellato per mitigare la possibile discriminazione presente.
Nel campo del in-processing sono stati sfruttati le due soluzioni descritte in precedenza fornite dalle due librerie. I risultati ottenuti dai modelli in in-processing verranno, poi, incrociati con i risultati dei modelli *standard* e dei modelli *fair*.
In fase di post-processing, vengono riutilizzati i modelli *Standard* realizzati andando ad effettuare le dovute valutazioni e modifiche necessarie alle predizioni fornite dai modelli.
Concludendo, possiamo evidenziare come gli scripts su cui vengono usate le due librerie di fairness generino **12** modelli in totale per libreria. Infine, vengono valutate le predizioni dei modelli *standard* e dei modelli *fair* per valuare i risultati di Fairness ottenuti.

[8] `https://github.com/ImCirot/Progetto-Tesi`

Al termine della fase di training dei modelli, viene generato un report finale in cui tramite la funzioni offerte dai modelli, viene richiesto al modello di predire l'intero dataset di input e restituire l'accuracy, f1-score, precision e recall ottenuti. Questi risultati sono poi di vitale importanza per poter confrontare i modelli ottenuti sul loro comportamento generale.

Per il dataset *UTKFace*, come già descritto in precedenza, vengono sviluppati 2 modelli *standard* tramite gli algoritmi *ResNet50* e *MobileNetV2*. In particolare, vengono sfruttare tutte le soluzioni di pre, in e post-processing utilizzate per i modelli di ML, generando quindi 2 diverse reti per ogni soluzione di pre, in e post-processing.

Questi modelli ci permetteranno non solo di fare dei confronti sui risultati ottenuti da un modello **standard** ed un modello **fair**, ma anche importanti informazioni sui consumi necessari a sviluppare modelli più complessi e pesanti.

Per concludere questa sezione, in **Figura 3.1**, con le operazioni di pre-in-post-processing, l'addestramento e testing dei modelli, i report finali prodotti ed i modelli ottenuti salvati, pronti per essere riutilizzati in ulteriori contesti.

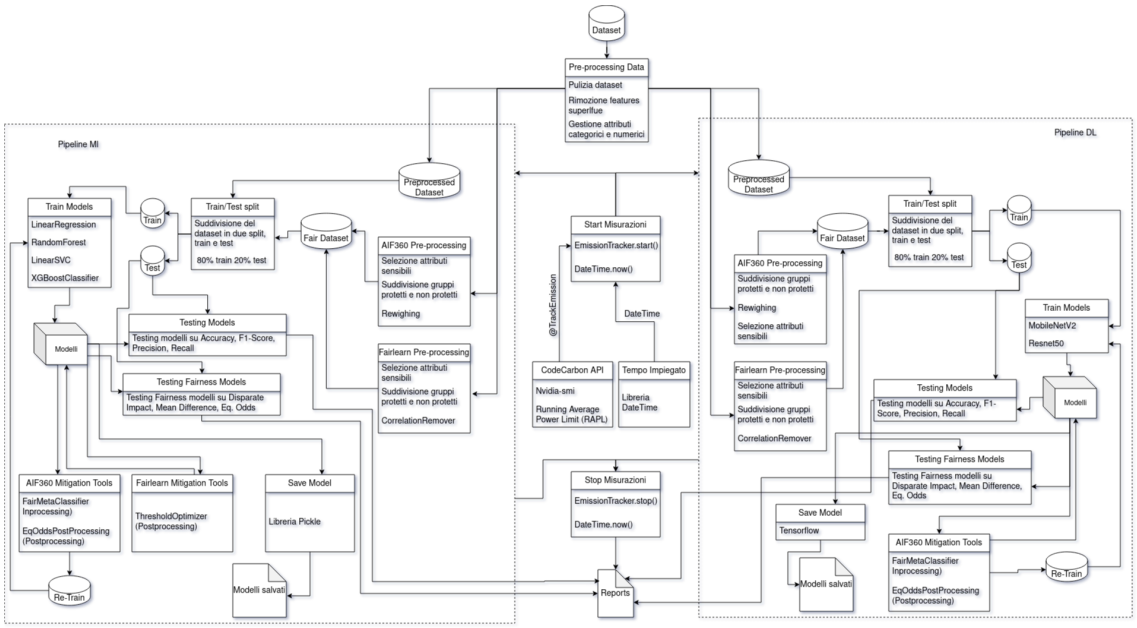

Figura 3.1: La pipeline di produzione dell'intero studio.

CAPITOLO 4

Analisi dei risultati

In questa sezione vengono presentati e analizzati i risultati ottenuti, al fine di fornire delle risposte quanto più dettagliate per le RQs proposte.
In particolare, verranno presentati i risultati analitici ottenuti dai reports generati durante l'esecuzione del codice per poter poi fornire delle stime valutative sul comportamento dei modelli in contesti di discriminazione, equità, sostenibilità e qualità dei modelli stessi.
Di seguito, vengono presentate nuovamente le RQs e fornite delle risposte sulla base delle considerazioni pertinenti sui risultati ottenuti.

4.1 RQ1: Che impatto hanno gli strumenti di Fairness sulla sostenibilità sociale del prodotto?

Con questa RQ si vuole valutare l'impatto che gli strumenti di Fairness e le soluzioni proposte per le fasi di pre, in e post-processing effettuate hanno a livello di equità del prodotto realizzato. Per poter fornire delle dovute riflessioni e conclusioni sulle strategie, andiamo a considerare i risultati ottenuti dalle due librerie confrontando singolarmente i tipi di operazione, presentando i risultati ottenuti, in termini di **Disparate Impact**, **Mean Difference** e **Equalized Odds** dai vari modelli ottenuti da diverse soluzioni, distinguendo anche i risultati ottenuti in termini di Fairness per i diversi gruppi protetti evidenziati, in particolare, verranno presentati i valori delle metriche di Fairness valutate singolarmente per i singoli sottogruppi protetti, andando a fornire una visione anche più specifica per ogni sottogruppo, evidenziando eventuali discrepanze nei risultati ottenuti in base al tipo di attributo protetto.
Presentiamo prima, i risultati ottenuti in termini di Fairness sia sui datasets standard, che sui datasets modificati che quelli prodotti dai modelli, andando a arricchire la risposta esaminando opportunamente tutte le possibili soluzioni proposte dalle due librerie per ogni tipo di operazione.
Partendo dal pre-processing, nelle **Tabelle 4.1, 4.2, 4.3 e 4.4** è possibile visualizzare le metriche di Fairness calcolate prima e dopo le operazioni effettuate per ogni dataset dello studio, andando ad evidenziare eventuali miglioramenti e peggioramenti, inoltre, per semplificare la lettura e la stesura delle informazioni, verranno utilizzati gli acronimi **AIF** e **FLN** per indicare le due librerie di Fairness, AIFairness360 e Fairlearn rispettivamente; per gli algoritmi verranno usati gli acronimi **LR** (Logistic Regression), **RF** (Random Forest), **SVM** (Support Vector Machine) e **XGB** (XGBoost).

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.20	0.36	0.16	0.0 (100%)↓	1.0 (177.78%)↑	0.16
LR	AIF	age	-0.21	0.42	-0.14	-0.20 (0.48%)↓	0.44 (4.76%)↑	-0.14
RF	AIF	sex	-0.20	0.36	-0.07	0.0 (100%)↓	1.0 (177.78%)↑	-0.07
RF	AIF	age	-0.21	0.42	-0.13	-0.20 (0.48%)↓	0.44 (4.76%)↑	-0.13
SVM	AIF	sex	-0.20	0.36	0.19	0.0 (100%)↓	1.0 (177.78%)↑	0.19
SVM	AIF	age	-0.21	0.42	-0.14	-0.20 (0.48%)↓	0.44 (4.76%)↑	-0.14
XGB	AIF	sex	-0.20	0.36	0.14	0.0 (100%)↓	1.0 (177.78%)↑	0.14
XGB	AIF	age	-0.21	0.42	-0.09	-0.20 (0.48%)↓	0.44 (4.76%)↑	-0.09
Oggetto utilizzato: **Reweighing**								
LR	FLN	sex	0.18	0.31	0.09	0.41 (127.78%)↑	0.48 (54.84%)↑	0.40 (344.44%)↑
LR	FLN	age	0.23	0.31	0.17	0.43 (86.96%)↑	0.52 (67.74%)↑	0.46 (170.59%)↑
RF	FLN	sex	0.18	0.35	0.08	0.10 (44.44%)↓	0.41 (17.14%)↑	0.04 (50%)↓
RF	FLN	age	0.21	0.37	0.12	0.11 (47.62%)↓	0.42 (13.51%)↑	0.03 (75%)↓
SVM	FLN	sex	0.18	0.32	0.08	0.41 (127.78%)↑	0.50 (56.25%)↑	0.40 (400%)↑
SVM	FLN	age	0.22	0.31	0.16	0.41 (86.36%)↑	0.54 (74.19%)↑	0.44 (175%)↑
XGB	FLN	sex	0.17	0.37	0.06	0.20 (17.65%)↑	0.53 (43.24%)↑	0.14 (133.33%)↑
XGB	FLN	age	0.21	0.36	0.12	0.20 (4.76%)↓	0.58 (61.11%)↑	0.16 (33.33%)↑
Oggetto utilizzato: **CorrelationRemover**								

Tabella 4.1: Le metriche di Fairness calcolate prima e dopo le operazioni di pre-processing sull'Adult Dataset.

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.07	0.90	-0.12	0.01 (114.29%)↓	1.01 (12.22%)↑	-0.0 (100%) ↓
LR	AIF	age	-0.09	0.88	-0.14	-0.01 (88.89%)↓	0.98 (11.36%)↑	-0.06 (57.14%)↓
RF	AIF	sex	-0.07	0.90	-0.02	0.01 (114.29%)↓	1.01 (12.22%)↑	-0.03 (50%)↑
RF	AIF	age	-0.09	0.88	-0.01	-0.01 (88.89%)↓	0.98 (11.36%)↑	0.01
SVM	AIF	sex	-0.07	0.90	-0.12	0.01 (114.29%)↓	1.01 (12.22%)↑	-0.03 (75%)↓
SVM	AIF	age	-0.09	0.88	-0.11	-0.01 (88.89%)↓	0.98 (11.36%)↑	-0.06 (45.45%)↓
XGB	AIF	sex	-0.07	0.90	-0.04	0.01 (114.29%)↓	1.01 (12.22%)↑	-0.01 (75%)↓
XGB	AIF	age	-0.09	0.88	-0.05	-0.01 (88.89%)↓	0.98 (11.36%)↑	0.00 (100%)↓
Oggetto utilizzato: **Reweighing**								
LR	FLN	sex	0.18	0.65	0.25	0.06 (66.67%)↑	0.0 (100%)↓	0.07 (72%)↑
LR	FLN	age	0.18	0.67	0.17	0.00 (100%)↑	0.75 (11.94%)↑	0.00 (100%)↑
RF	FLN	sex	0.16	0.84	0.38	0.01 (93.75%)↑	0.99 (17.86%)↑	0.05 (86.84%)↑
RF	FLN	age	0.10	0.90	0.29	0.02 (80%)↑	0.98 (8.89%)↑	0.06 (79.31%)↑
SVM	FLN	sex	0.18	0.65	0.25	0.06 (66.67%)↑	0.0 (100%)↓	0.07 (72%)↑
SVM	FLN	age	0.16	0.71	0.17	0.00 (100%)↑	0.75 (5.63%)↑	0.00 (100%)↑
XGB	FLN	sex	0.34	0.66	0.63	0.50 (47.06%)↓	0.0 (100%)↓	0.57 (9.52%)↑
XGB	FLN	age	0.15	0.83	0.39	0.14 (7.14%)↑	0.72 (13.25%)↓	0.15 (61.54%)↑
Oggetto utilizzato: **CorrelationRemover**								

Tabella 4.2: Le metriche di Fairness calcolate prima e dopo le operazioni di pre-processing sul German Credit Dataset.

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.30	0.47	0.05	-0.01 (96.67%)↓	0.99 (110.64%)↑	0.20 (300%)↑
LR	AIF	age	-0.29	0.51	-0.26	-0.02 (93.10%)↓	0.95 86.27(%)↑	-0.10 (61.54%)↓
RF	AIF	sex	-0.30	0.47	0.13	-0.01 (96.67%)↓	0.99 (110.64%)↑	0.09 (30.77%)↓
RF	AIF	age	-0.29	0.51	-0.34	-0.02 (93.10%)↓	0.95 (86.27%)↑	-0.05 (85.29%)↓
SVM	AIF	sex	-0.30	0.47	0.05	-0.01 (96.67%)↓	0.99 (110.64%)↑	0.20 (300%)↑
SVM	AIF	age	-0.29	0.51	-0.26	-0.02 (93.10%)↓	0.95 (86.27%)↑	-0.10 (61.54%)↓
XGB	AIF	sex	-0.30	0.47	-0.02	-0.01 (96.67%)↓	0.99 (110.64%)↑	0.02
XGB	AIF	age	-0.29	0.51	-0.14	-0.02 (93.10%)↓	0.95 (86.27%)↑	-0.10 (28.57%)↓
Oggetto utilizzato: **Reweighing**								
LR	FLN	sex	0.27	0.58	0.12	0.56 (107.41%)↑	0.25 (56.90%)↓	0.50 (316.68%)↑
LR	FLN	age	0.37	0.43	0.29	0.37	0.43	0.35 (20.69%)↑
RF	FLN	sex	0.22	0.63	0.13	0.24 (9.10%)↑	0.60 (4.76%)↓	0.12 (7.69%)↓
RF	FLN	age	0.40	0.36	0.34	0.34 (15%)↓	0.45 (25%)↑	0.29 (14.71%)↓
SVM	FLN	sex	0.29	0.53	0.10	0.63 (117.24%)↑	0.18 (66.04%)↓	0.50 (400%)↑
SVM	FLN	age	0.32	0.47	0.26	0.21 (34.38%)↓	0.65 (38.30%)↑	0.29 (11.54%)↑
XGB	FLN	sex	0.22	0.65	0.15	0.07 (68.18%)↓	0.86 (32.31%)↑	0.17 (13.33%)↑
XGB	FLN	age	0.39	0.42	0.53	0.40 (2.56%)↑	0.36 (14.29%)↓	0.41 (22.64%)↓
Oggetto utilizzato: **CorrelationRemover**								

Tabella 4.3: Le metriche di Fairness calcolate prima e dopo le operazioni di pre-processing sul Heart Disease Dataset.

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.10	0.82	-0.22	-0.10	0.81 (1.22%)↓	-0.22
LR	AIF	age	0.11	1.26	0.27	0.11	1.25 (0.79%)↑	0.26 (3.70%)↓
RF	AIF	sex	-0.10	0.82	-0.01	-0.10	0.81 (1.22%)↓	-0.02 (100%)↑
RF	AIF	age	0.11	1.26	0.01	0.11	1.25 (0.79%)↑	0.05 (400%)↑
SVM	AIF	sex	-0.10	0.82	-0.22	-0.10	0.81 (1.22%)↓	-0.22
SVM	AIF	age	0.11	1.26	0.27	0.11	1.25 (0.79%)↑	0.26 (3.70%)↓
XGB	AIF	sex	-0.10	0.82	-0.10	-0.10	0.81 (1.22%)↓	-0.11 (10%)↑
XGB	AIF	age	0.11	1.26	0.14	0.11	1.25 (0.79%)↑	0.15 (7.14%)↑
Oggetto utilizzato: **Reweighing**								
LR	FLN	sex	0.25	0.62	0.22	0.28 (12%)↑	0.58 (6.45%)↓	0.26 (18.18%)↑
LR	FLN	age	0.25	0.58	0.25	0.35 (40%)↑	0.46 (20.69%)↓	0.34 (36%)↑
RF	FLN	sex	0.13	0.78	0.09	0.01 (92.31%)↓	0.97 (24.36%)↑	0.07 (22.22%)↓
RF	FLN	age	0.17	0.72	0.12	0.17	0.53 (26.39%)↑	0.23 91.(91.67%)↑
SVM	FLN	sex	0.24	0.62	0.22	0.38 (58.33%)↑	0.47 (24.19%)↓	0.38 (72.73%)↑
SVM	FLN	age	0.26	0.57	0.25	0.44 (69.23%)↑	0.35 (38.60%)↓	0.44 (76%)↑
XGB	FLN	sex	0.17	0.73	0.15	0.08 (52.94%)↓	0.83 (13.70%)↑	0.06 (60%)↓
XGB	FLN	age	0.20	0.68	0.16	0.10 (50%)↓	0.79 (16.18%)↑	0.09 (43.75%)↓
Oggetto utilizzato: **CorrelationRemover**								

Tabella 4.4: Le metriche di Fairness calcolate prima e dopo le operazioni di pre-processing sul Home Credit Default Risk Dataset.

Modello	Lib.	M.Diff.	DI	EqOdds	M. Diff	DI	EqOdds
Resnet50	AIF	-0.36	0.93	0.03	0.0(100%)↓	1.0(100%)↓	-0.00(100%)↓
MobNetV2	AIF	-0.04	0.93	-0.01	0.0(100%)↓	1.0(100%)↓	-0.04(300%)↑
Oggetto utilizzato: **Reweighing**							
Resnet50	FLN	0.04	0.91	0.09	0.02(50%)↓	0.96(5.49%)↓	0.07(22.22%)↓
MobNetV2	FLN	0.03	0.92	0.08	0.05(66.67%)↑	0.84(8.70%)↑	0.08
Oggetto utilizzato: **CorrelationRemover**							

Tabella 4.5: Le metriche di Fairness calcolate prima e dopo le operazioni di pre-processing sull'UTKFace Dataset (DL).

In fase di pre-processing, possiamo notare come la strategia di pre-processing offerta dalla libreria AIF360 ottenga, in generale, i risultati migliori, in quanto, in tutti i datasets esplorati è riuscita a migliorare il valore di *Disparate Impact* e *Mean Difference*, invece il valore *Eq. Odds* rimane spesso invariato, migliora o peggiora di circa 0.01 poiché, in generale, questa metrica di Fairness è molto più impattante in situazioni di in e post-processing, in quanto viene calcolata utilizzando le predizioni del modello piuttosto che il dataset, che di solito, viene modificato in pre-processing. Invece, la libreria Fairlearn, ottiene risultati molto altalenanti, addirittura in molti casi peggiora di molto i valori delle metriche, riuscendo poche volte a realmente migliorare i valori. Questa situazione è sicuramente dovuta alla strategia e al modo in cui vengono calcolate le metriche, in quanto anche se stiamo trattando pre-processing, quindi modifiche al solo dataset e non ai modelli, per calcolare le metriche, Fairlearn utilizza anche le predizioni del modello. Questa cosa può condizionare, com'è possibile vedere, i risultati. Essendo l'unica strategia di mitigazione di Fairness di pre-processing, purtroppo, i risultati ottenuti sono spesso inconcludenti o peggiori della controparte base e dei risultati dell'altra libreria.

Possiamo concludere che, confrontando le due librerie, la soluzione offerta da AIF360 sia da preferire e, in linea di massima, molto generalizzabile, in quanto i risultati prodotti su diversi datasets sembrano, a prescindere dai valori rimanenti, riuscire a mitigare i valori e produrre modelli molto più equi della controparte dai Fairlearn.

Passando ora alla fase di in-processing, per via della mancanza di una vera soluzione di in-processing da parte della libreria Fairlearn, vengono presentati i risultati ottenuti in in-processing dalla sola libreria AIF360 nelle **Tabelle 4.6, 4.8, 4.7, 4.9 e 4.10**, sfruttando anche qui gli acronimi **LR** (Logistic Regression), **RF** (Random Forest), **SVM** (Support Vector Machine) e **XGB** (XGBoost) per indicare facilmente i diversi modelli.

Modello	Attrb.	M.Diff.	DI	EqOdds	M. Diff	DI	EqOdds
LR	sex	-0.20	0.36	-0.09	0.07 (135%)↓	1.12 (211.11%)↑	0.01 (111.11%)↓
LR	age	-0.21	0.43	-0.17	-0.01 (95%)↓	0.99 (130.23%)↑	0.02 (111.77%)↓
RF	sex	-0.20	0.36	-0.05	0.07 (135%)↓	1.12 (211.11%)↑	0.02 (140%)↓
RF	age	-0.21	0.43	-0.12	-0.01 (95%)↓	0.99 (130.23%)↑	0.02 (116.67%)↓
SVM	sex	-0.20	0.36	-0.08	0.07 (135%)↓	1.12 (211.11%)↑	0.02 (125%)↓
SVM	age	-0.21	0.43	-0.16	-0.01 (95%)↓	0.99 (130.23%)↑	0.02 (112.5%)↓
XGB	sex	-0.20	0.36	-0.03	0.07 (135%)↓	1.12 (211.11%)↑	0.02 (166.67%)↓
XGB	age	-0.21	0.43	-0.12	-0.01 (95%)↓	0.99 (130.23%)↑	0.01 (108.33%)↓
Oggetto utilizzato: **MetaFairClassifier**							

Tabella 4.6: Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Modello	M.Diff.	DI	EqOdds	M. Diff	DI	EqOdds
Resnet50	-0.04	0.93	0.09	-0.0 (100%)↓	1.14 (22.58%)↑	0.02 (77.78%)↓
MobNetV2	-0.04	0.93	0.09	0.02 (150%)↓	1.14 (22.58%)↑	-0.02 (122.22%)↓
Oggetto utilizzato: **MetaFairClassifier**						

Tabella 4.7: Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sull'UTKFace Dataset (DL).

Modello	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	sex	-0.07	0.91	-0.12	-0.12 (71.43%)↑	0.86 (5.50%)↓	-0.06 (50%)↓
LR	age	-0.11	0.88	-0.14	-0.07 (36.36%)↓	0.92 (4.55%)↑	-0.05 (64.29%)↓
RF	sex	-0.07	0.91	-0.02	-0.12 (71.43%)↑	0.86 (5.50%)↓	-0.01 (50%)↓
RF	age	-0.11	0.88	-0.01	-0.07 (36.36%)↓	0.92 (4.55%)↑	-0.02 (100%)↑
SVM	sex	-0.07	0.91	-0.12	-0.12 (71.43%)↑	0.86 (5.50%)↓	-0.07 (41.67%)↓
SVM	age	-0.11	0.88	-0.11	-0.07 (36.36%)↓	0.92 (4.55%)↑	-0.04 (63.63%)↓
XGB	sex	-0.07	0.91	-0.04	-0.12 (71.43%)↑	0.86 (5.50%)↓	-0.01 (75%)↓
XGB	age	-0.11	0.88	-0.05	-0.07 (36.36%)↓	0.92 (4.55%)↑	-0.04 (80%)↓
Oggetto utilizzato: **MetaFairClassifier**							

Tabella 4.8: Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Modello	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	sex	-0.28	0.46	0.05	-0.33 (17.86%)↑	0.35 (23.91%)↓	0.05
LR	age	-0.32	0.47	-0.26	-0.31 (3.13%)↓	0.44 (6.38%)↓	-0.26
RF	sex	-0.28	0.46	0.13	-0.33 (17.86%)↑	0.35 (23.91%)↓	0.09 (30.77%)↓
RF	age	-0.32	0.47	-0.34	-0.31 (3.13%)↓	0.44 (6.38%)↓	-0.22 (35.29%)↓
SVM	sex	-0.28	0.46	0.05	-0.33 (17.86%)↑	0.35 (23.91%)↓	0.05
SVM	age	-0.32	0.47	-0.26	-0.31 (3.13%)↓	0.44 (6.38%)↓	-0.26
XGB	sex	-0.28	0.46	-0.02	-0.33 (17.86%)↑	0.35 (23.91%)↓	0.09 (550%)↓
XGB	age	-0.32	0.47	-0.14	-0.31 (3.13%)↓	0.44 (6.38%)↓	-0.22 (57.14%)↓
Oggetto utilizzato: **MetaFairClassifier**							

Tabella 4.9: Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Modello	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	sex	-0.11	0.82	-0.22	$0.0_{(100\%)\uparrow}$	$1.0_{(100\%)\uparrow}$	-0.22
LR	age	0.11	1.26	0.25	$0.0_{(100\%)\downarrow}$	$1.0_{(100\%)\downarrow}$	0.25
RF	sex	-0.11	0.82	-0.04	$0.0_{(100\%)\uparrow}$	$1.0_{(100\%)\uparrow}$	-0.04
RF	age	0.11	1.26	0.04	$0.0_{(100\%)\downarrow}$	$1.0_{(100\%)\downarrow}$	0.04
SVM	sex	-0.11	0.82	-0.22	$0.0_{(100\%)\uparrow}$	$1.0_{(100\%)\uparrow}$	-0.22
SVM	age	0.11	1.26	0.25	$0.0_{(100\%)\downarrow}$	$1.0_{(100\%)\downarrow}$	0.25
XGB	sex	-0.11	0.82	-0.12	$0.0_{(100\%)\uparrow}$	$1.0_{(100\%)\uparrow}$	-0.12
XGB	age	0.11	1.26	0.13	$0.0_{(100\%)\downarrow}$	$1.0_{(100\%)\downarrow}$	0.13
Oggetto utilizzato: **MetaFairClassifier**							

Tabella 4.10: Le metriche di Fairness calcolate prima e dopo le operazioni di in-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

In fase di in-processing possiamo notare come i risultati evidenziati siano altalenanti, in generale, però, i risultati sono positivi, andando nella maggior parte dei casi a migliorare le metriche in generale. L'unica eccezione sono i modelli realizzati in in-processing sul Heart Disease Dataset, che, dopo la fase di in-processing, ottengono *Disparate Impact* molto più bassi rispetto agli originali, già ampiamente non soddisfacenti. Questo ci permette di dire che la strategia proposta da AIF360 per la fase di in-processing è più che valida, ma bisognerebbe approfondire maggiormente il discorso, presentando altri dataset dove le metriche peggiorano, per poter definire formalmente come mai questa strategia non ha funzionato su dataset come l'Heart Disease Dataset.

Infine, in fase di post-processing, nelle **Tabelle 4.11, 4.12, 4.13, 4.14 e 4.15**, vengono presentati i risultati ottenuti dalle due strategie di post-processing scelte dalle librerie AIF360 e Fairlearn, evidenziato però, come non sia stato possibile sfruttare la soluzione offerta da Fairlearn sui modelli di DL, in quanto incompatibile con i modelli di Tensorflow.

Per facilitare la lettura e stesura dei risultati, anche qui, sfruttiamo gli acronimi **AIF** e **FLN** per le due librerie AIFairness360 e Fairlearn e **LR** (Logistic Regression), **RF** (Random Forest), **SVM** (Support Vector Machine) e **XGB** (XGBoost) per indicare i modelli.

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.20	0.38	-0.09	-0.10 (50%)↓	0.60 (57.90%)↑	-0.08 (11.11%)↓
LR	AIF	age	-0.20	0.37	-0.17	-0.16 (%)↓	0.45 (21.62%)↑	0.06 (135.29%)↓
RF	AIF	sex	-0.20	0.38	-0.05	-0.10 (50%)↓	0.63 (65.79%)↑	-0.04 (20%)↓
RF	AIF	age	-0.20	0.42	-0.12	-0.16 (20%)↓	0.45 (40.63%)↑	-0.04 (66.67%)↓
SVM	AIF	sex	-0.20	0.38	-0.08	-0.10 (50%)↓	0.60 (57.90%)↑	-0.07 (12.5%)↓
SVM	AIF	age	-0.20	0.36	-0.16	-0.16 (20%)↓	0.42 (16.67%)↑	-0.07 (56.25%)↓
XGB	AIF	sex	-0.20	0.38	-0.03	-0.11 (45%)↓	0.58 (52.63%)↑	-0.05 (66.67%)↓
XGB	AIF	age	-0.20	0.40	-0.12	-0.17 (15%)↓	0.41 (2.5%)↑	-0.04 (66.67%)↓
Oggetto utilizzato: **EqOddsPostprocessing**								
LR	FLN	sex	0.18	0.31	0.09	0.01 (94.44%)↑	0.95 (206.45%)↑	0.35 (288.89%)↓
LR	FLN	age	0.23	0.31	0.17	0.02 (91.30%)↑	0.91 (193.55%)↑	0.23 (35.29%)↓
RF	FLN	sex	0.18	0.35	0.08	0.08 (55.56%)↑	0.64 (82.86%)↑	0.38 (375%)↓
RF	FLN	age	0.21	0.37	0.12	0.03 (85.71%)↑	0.86 (132.43%)↑	0.26 (116.67%)↓
SVM	FLN	sex	0.18	0.32	0.08	0.01 (94.44%)↑	0.94 (193.75%)↑	0.35 (337.5%)↓
SVM	FLN	age	0.22	0.31	0.16	0.02 (90.91%)↑	0.91 (193.55%)↑	0.22 (37.5%)↓
XGB	FLN	sex	0.17	0.37	0.06	0.01 (94.12%)↑	0.93 (151.35%)↑	0.35 (483.33%)↓
XGB	FLN	age	0.21	0.36	0.12	0.01 (95.24%)↑	0.94 (161.11%)↑	0.25 (108.33%)↓
Oggetto utilizzato: **ThresholdOptimizer**								

Tabella 4.11: Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.12	0.86	-0.12	-0.05(58.33%)↓	0.88(2.33%)↑	-0.13(8.33%)↑
LR	AIF	age	-0.15	0.70	-0.14	-0.13(13.33%)↓	0.75(7.14%)↑	-0.10(28.57%)↓
RF	AIF	sex	-0.12	0.86	-0.02	-0.04(66.67%)↓	0.96(11.63%)↑	-0.00(100%)↓
RF	AIF	age	-0.09	0.91	-0.01	-0.07(22.22%)↓	0.93(2.20%)↑	0.00(100%)↓
SVM	AIF	sex	-0.12	0.86	-0.12	-0.05(58.33%)↓	0.88(2.33%)↑	-0.10(16.67%)↓
SVM	AIF	age	-0.11	0.77	-0.11	-0.09(22.22%)↓	0.82(6.49%)↑	-0.05(54.55%)↓
XGB	AIF	sex	-0.12	0.86	-0.04	-0.03(75%)↓	0.97(12.79%)↑	-0.04
XGB	AIF	age	-0.15	0.84	-0.05	-0.14(6.67%)↓	0.84	-0.05
Oggetto utilizzato: **EqOddsPostprocessing**								
LR	FLN	sex	0.18	0.65	0.25	0.24(33.33%)↑	0.76(16.92%)↑	0.50(100%)↑
LR	FLN	age	0.18	0.67	0.17	0.04(77.78%)↓	0.95(41.79%)↑	0.12(29.41%)↓
RF	FLN	sex	0.16	0.84	0.38	0.26(62.5%)↑	0.74(11.91%)↓	0.50(31.58%)↑
RF	FLN	age	0.10	0.90	0.29	0.12(20%)↑	0.87(3.33%)↓	0.32(10.35%)↑
SVM	FLN	sex	0.18	0.65	0.25	0.27(50%)↑	0.71(9.23%)↑	0.50(100%)↑
SVM	FLN	age	0.16	0.71	0.17	0.04(75%)↓	0.96(35.21%)↑	0.03(82.35%)↓
XGB	FLN	sex	0.34	0.66	0.63	0.31(8.82%)↓	0.69(4.55%)↑	0.50(20.64%)↓
XGB	FLN	age	0.15	0.83	0.39	0.07(53.33%)↓	0.92(10.84%)↑	0.27(30.77%)↓
Oggetto utilizzato: **ThresholdOptimizer**								

Tabella 4.12: Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.37	0.47	0.05	-0.18(51.35%)↓	0.72(53.19%)↑	0.05
LR	AIF	age	-0.41	0.40	-0.26	-0.37(9.76%)↓	0.43(7.5%)↑	-0.26
RF	AIF	sex	-0.37	0.47	0.13	-0.17(54.05%)↓	0.71(51.06%)↑	0.13
RF	AIF	age	-0.42	0.35	-0.34	-0.40(4.76%)↓	0.36(2.86%)↑	-0.34
SVM	AIF	sex	-0.37	0.47	0.05	-0.24(35.14%)↓	0.60(27.66%)↑	0.05
SVM	AIF	age	-0.34	0.45	-0.26	-0.32(5.88%)↓	0.47(4.44%)↑	-0.26
XGB	AIF	sex	-0.37	0.47	-0.02	-0.13(64.87%)↓	0.79(68.09%)↑	-0.02
XGB	AIF	age	-0.44	0.39	-0.14	-0.39(11.37%)↓	0.42(7.69%)↑	-0.14
Oggetto utilizzato: **EqOddsPostprocessing**								
LR	FLN	sex	0.27	0.58	0.12	0.10(62.96%)↓	0.82(41.38%)↑	0.28(133.33%)↑
LR	FLN	age	0.37	0.43	0.29	0.14(62.16%)↓	0.74(72.09%)↑	0.31(6.90%)↑
RF	FLN	sex	0.22	0.63	0.13	0.07(68.18%)↓	0.80(26.98%)↑	0.25(92.31%)↑
RF	FLN	age	0.40	0.36	0.34	0.06(85%)↓	0.83(130.56%)↑	0.04(88.24%)↓
SVM	FLN	sex	0.29	0.53	0.10	0.16(44.83%)↓	0.68(28.30%)↑	0.43(330%)↑
SVM	FLN	age	0.32	0.47	0.26	0.10(68.75%)↓	0.78(65.96%)↑	0.25(3.85%)↓
XGB	FLN	sex	0.22	0.65	0.15	0.03(86.37%)↓	0.93(43.08%)↑	0.21(40%)↑
XGB	FLN	age	0.39	0.42	0.53	0.02(94.87%)↓	0.96(128.57%)↑	0.11(79.25%)↓
Oggetto utilizzato: **ThresholdOptimizer**								

Tabella 4.13: Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Mod.	Lib.	Attr.	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
LR	AIF	sex	-0.09	0.84	-0.22	-0.02(77.78%)↓	0.96(14.29%)↑	-0.19(13.64%)↓
LR	AIF	age	0.18	1.46	0.25	0.13(27.78%)↓	1.32(9.58%)↓	0.11(56%)↓
RF	AIF	sex	-0.09	0.84	-0.04	-0.06(33.33%)↓	0.90(7.14%)↑	-0.19(375%)↑
RF	AIF	age	0.14	1.31	0.04	0.13(7.14%)↓	1.30(0.76%)↓	0.11(175%)↑
SVM	AIF	sex	-0.09	0.84	-0.22	-0.02(77.78%)↓	0.96(14.29%)↑	-0.19(13.64%)↓
SVM	AIF	age	0.19	1.46	0.25	0.13(31.58%)↓	1.32(9.59%)↓	0.11(56%)↓
XGB	AIF	sex	-0.09	0.84	-0.12	-0.04(55.56%)↓	0.94(11.91%)↑	-0.19(58.33%)↑
XGB	AIF	age	0.16	1.35	0.13	0.13(18.75%)↓	1.30(3.70%)↓	0.11(15.39%)↓
Oggetto utilizzato: **EqOddsPostprocessing**								
LR	FLN	sex	0.25	0.62	0.22	0.00(100%)↓	1.0(100%)↑	0.04(81.82%)↓
LR	FLN	age	0.25	0.58	0.25	0.01(96%)↓	0.98(68.97%)↑	0.07(72%)↓
RF	FLN	sex	0.13	0.78	0.09	0.25(92.31%)↑	0.47(39.74%)↓	0.47(422.22%)↑
RF	FLN	age	0.17	0.72	0.12	0.24(41.18%)↑	0.68(5.56%)↓	0.48(300%)↑
SVM	FLN	sex	0.24	0.62	0.22	0.01(95.83%)↓	0.99(59.67%)↑	0.04(81.82%)↓
SVM	FLN	age	0.26	0.57	0.25	0.01(96.15%)↓	0.98(71.93%)↑	0.07(72%)↓
XGB	FLN	sex	0.17	0.73	0.15	0.01(94.12%)↓	0.99(35.62%)↑	0.06(60%)↓
XGB	FLN	age	0.20	0.68	0.16	0.01(95%)↓	0.98(44.12%)↑	0.08(50%)↓
Oggetto utilizzato: **ThresholdOptimizer**								

Tabella 4.14: Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, DI = Disparate Impact)

Modello	M.Diff.	DI	EqOdds	M.Diff.	DI	EqOdds
Resnet50	-0.05	0.91	0.02	0.55(1200%)↑	0.98(7.69%)↑	0.56(2700%)↑
MobNetV2	-0.05	0.91	0.01	-0.01(80%)↓	0.98(7.69%)↑	0.01
Oggetto utilizzato: **EqOddsPostprocessing**						

Tabella 4.15: Le metriche di Fairness calcolate prima e dopo le operazioni di post-processing sull'UTKFace Dataset (DL).

In fase di post-processing possiamo notare come la soluzione proposta dalla libreria AIF360, in tutti i datasets, vada a migliorare le metriche di Fairness individuate, con pochissimi ininfluenti casi in cui qualche metrica è peggiorata.
Lo stesso vale per la soluzione proposta da Fairlearn, in particolare però, possiamo notare come molto frequentemente ci sia un peggioramento delle metrica *Equalized Odds* su almeno uno dei quattro modelli, con il caso più eclatante dell'Adult Dataset dove le soluzioni proposte da Fairlean peggiorino sistematicamente il valore di *Equalized Odds* ottenuto.
Per concludere questa RQ, forniamo la seguente risposta:

👉 **Answer to RQ$_1$.** La libreria **AIFairness360 (AIF360)** ha ottenuto, in tutti e tre gli ambiti di pre, in e post-processing, i risultati migliori in termini di metriche di Fairness con le sue soluzioni proposte.
Risulta essere più completa dal punto di vista delle soluzioni proposte, spesso con soluzioni generalizzabili che si adattano discretamente ai diversi datasets che sono stati proposti rendendo la libreria utilizzabile senza particolari studi sulle caratteristiche del dataset, rispetto alla libreria **Fairlearn**, che con le sue soluzioni, spesso ha riportato problemi su determinati datasets ottenendo anche risultati "peggiori" dal punto di vista Fairness. Questi fallimenti richiederebbero ulteriori studi sia sulle caratteristiche dei datasets dove la libreria ha fallito, sia sul comportamento delle soluzioni proposte per poter adattare il datasets alle operazioni. Entrambe però rimangono delle soluzioni più che valide per poter mitigare problemi di Fairness.

4.2 RQ2: Che impatto hanno gli strumenti di Fairness sulla sostenibilità ambientale del prodotto?

In questa RQ andremo ad esaminare i risultati prodotti in termini di consumo e utilizzo di risorse dai vari scripts per addestrare e validare i diversi modelli proposti sui vari datasets individuati nello studio.
Questa RQ è molto importante in termini di sostenibilità, nasce dall'idea di fornire una visione più completa del fenomeno dei consumi legati al mondo dell'IA, non solo nel mondo del DL ma a qualunque modello di ML o DL che computazionalmente ha bisogni importati.
Per poter rispondere a questa domanda è stata utilizzata la libreria **CodeCarbon**, andando a produrre un report di consumo e richiesta risorse ad ogni esecuzione di script. In particolare, gli scripts di modelli ML sono stati reiterati 10 volte per poter produrre dei risultati concreti ed eventualmente stabili per poter fornire delle risposte quanto più generali e accurate possibili sui consumi richiesti dai vari modelli su datasets di diverse dimensioni, sia in grandezza che in numero di features.
In totale sono state effettuate oltre *200* misurazioni.
In particolare, nei risultati prodotti, in particolare per la fase di DL, troviamo valori molto più grandi, parlando di circa **5g** di CO2 prodotta per una singola esecuzione di uno script, che addestra un solo modello. Questi grafici ci permettono di vedere subito la differenza di tempo e di risorse necessaria richiesta dall'addestramento e testing di un modello di DL rispetto ad un modello di ML. Per poter giungere a delle conclusioni e fornire un quadro più dettagliato dei consumi, in **Tabella 4.16 e 4.17** troviamo i consumi dei modelli standard rispettivamente di ML e di DL sui diversi datasets; nelle **Tabelle 4.18, 4.19, 4.20 e 4.21**, troviamo i consumi, in ordine di operazione di pre-in-post-processing rispettivamente, dei modelli di ML e DL, prima per la libreria **AIFairness360** e poi per la libreria **Fairlearn**, inoltre, evidenziati i miglioramenti o peggioramenti dei diversi valori di consumo rispetto ai modelli di ML e DL standard. Per semplificare e migliorare la leggibilità delle tabelle sono stati utilizzati diversi acronimi, **Dat.** (Dataset), **Cons. (Wh)** (Consumo energetico in Wh), **CO2 (g)** (CO2 prodotta in grammi), **T (s)** (Tempo richiesto in secondi).

Dataset	Cons. (Wh)	CO2 (g)	T (s)
Adult	0.57	0.20	36.19
German	0.08	0.03	3.27
Heart	0.04	0.001	1.60
Home	5.43	1.83	330.62

Tabella 4.16: I consumi dei modelli standard di ML sui diversi datasets.

Modello	Cons. (Wh)	CO2 (g)	T (s)
ResNet50	14.95	5.09	683.37
MobileNetV2	12.13	4.14	672.83

Tabella 4.17: I consumi dei modelli standard di DL sull'UTKFace dataset.

Dataset	Cons. (Wh)	CO2 (g)	T (s)	Cons. (Wh)	CO2 (g)	T (s)
Adult	0.57	0.20	36.19	0.68 (19.30%)↑	0.24 (20%)↑	42.07 (16.25%)↑
German	0.08	0.03	3.27	0.12 (50%)↑	0.04 (33.33%)↑	5.13 (56.88%)↑
Heart	0.04	0.001	1.60	0.06 (50%)↑	0.02 (1900%)↑	2.67 (66.88%)↑
Home	5.43	1.83	330.62	6.61 (21.73%)↑	2.13 (16.39%)↑	387.65 (17.25%)↑
Oggetto utilizzato: **Reweighing**						
Adult	0.57	0.20	36.19	1.47 (157.90%)↑	0.49 (145%)↑	90.84 (151.01%)↑
German	0.08	0.03	3.27	0.81 (912.5%)↑	0.27 (800%)↑	50.05 (1430.58%)↑
Heart	0.04	0.001	1.60	0.05 (25%)↑	0.02 (1900%)↑	2.85 (78.13%)↑
Home	5.43	1.83	330.62	6.45 (18.79%)↑	2.10 (14.75%)↑	381.74 (15.46%)↑
Oggetto utilizzato: **MetaFairClassifier**						
Adult	0.57	0.20	36.19	0.15 (73.68%)↓	0.05 (75%)↓	7.04 (78.06%)↓
German	0.08	0.03	3.27	0.06 (25%)↓	0.02 (33.33%)↓	2.48 (24.16%)↓
Heart	0.04	0.001	1.60	0.06 (50%)↑	0.001	1.74 (8.75%)↑
Home	5.43	1.83	330.62	0.38 (93%)↓	0.14 (92.35%)↓	22.77 (93.11%)↓
Oggetto utilizzato: **EqOddsPostprocessing**						

Tabella 4.18: I consumi dei modelli di ML, in ordine di operazione di pre-in-post-processing, per i vari datasets (AIF360).

Mod.	Cons. (Wh)	CO2 (g)	T (s)	Cons. (Wh)	CO2 (g)	T (s)
RN50	14.95	5.09	683.37	15.20 (1.67%)↑	5.18 (1.77%)↑	694.75 (1.67%)↑
MNV2	12.13	4.14	672.83	12.40 (2.23%)↑	4.23 (2.17%)↑	690.42 (2.61%)↑
Oggetto utilizzato: **Reweighing**						
RN50	14.95	5.09	683.37	14.87 (0.54%)↓	5.06 (0.59%)↓	687.76 (0.64%)↑
MNV2	12.13	4.14	672.83	12.47 (2.80%)↑	4.25 (2.66%)↑	691.90 (2.83%)↑
Oggetto utilizzato: **MetaFairClassifier**						
RN50	14.95	5.09	683.37	0.22 (98.53%)↓	0.07 (98.63%)↓	12.24 (98.21%)↓
MNV2	12.13	4.14	672.83	0.17 (98.60%)↓	0.06 (98.56%)↓	10.48 (98.44%)↓
Oggetto utilizzato: **EqOddsPostprocessing**						

Tabella 4.19: I consumi dei modelli di DL, in ordine di di pre-in-post-processing, sull'UTKFace dataset (AIF360). (RN50 = ResNet50, MNV2 = MobileNetV2)

Dataset	Cons. (Wh)	CO2 (g)	T (s)	Cons. (Wh)	CO2 (g)	T (s)
Adult	0.57	0.20	36.19	0.86(50.88%)↑	0.29(45%)↑	52.31(44.54%)↑
German	0.08	0.03	3.27	0.19(137.5%)↑	0.06(100%)↑	9.02(175.84%)↑
Heart	0.04	0.001	1.60	0.12(200%)↑	0.04(3900%)↑	6.33(295.63%)↑
Home	5.43	1.83	330.62	6.86(26.34%)↑	2.44(33.33%)↑	419.60(26.91%)↑
Oggetto utilizzato: **CorrelationRemover**						
Adult	-	-	-	-	-	-
German	-	-	-	-	-	-
Heart	-	-	-	-	-	-
Home	-	-	-	-	-	-
Nessun oggetto disponibile						
Adult	0.57	0.20	36.19	0.80(40.35%)↑	0.27(35%)↑	49.87(37.80%)↑
German	0.08	0.03	3.27	0.32(300%)↑	0.11(266.67%)↑	18.47(464.83%)↑
Heart	0.04	0.001	1.60	0.19(375%)↑	0.06(5900%)↑	10.35(546.88%)↑
Home	5.43	1.83	330.62	2.33(57.09%)↓	0.78(57.38%)↓	141.79(60.14%)↓
Oggetto utilizzato: **ThresholdOptimizer**						

Tabella 4.20: I consumi dei modelli di ML, in ordine di operazione di pre-in-post-processing, per i vari datasets (Fairlearn).

Mod.	Cons. (Wh)	CO2 (g)	T (s)	Cons. (Wh)	CO2 (g)	T (s)
RN50	14.95	5.09	683.37	15.48(3.55%)↑	5.28(3.73%)↑	697.30(2.01%)↑
MNV2	12.13	4.14	672.83	12.54(3.38%)↑	4.28(3.38%)↑	692.91(2.98%)↑
Oggetto utilizzato: **CorrelationRemover**						
RN50	-	-	-	-	-	-
MNV2	-	-	-	-	-	-
Nessun oggetto disponibile						
RN50	-	-	-	-	-	-
MNV2	-	-	-	-	-	-
Nessun oggetto disponibile						

Tabella 4.21: I consumi dei modelli di DL, in ordine di di pre-in-post-processing, sull'UTKFace dataset (Fairlearn). (RN50 = ResNet50, MNV2 = MobileNetV2)

Dalle Tabelle appena presente è subito possibile notare come i modelli di DL, tradizionalmente, richiedano il maggior numero di risorse e producano più CO2.
Le informazioni sui consumi ottenuti sia dai modelli di ML sia dai modelli di DL ci permettono di evidenziare come, in tempi brevi, comunque la quantità di risorsa energetica richiesta sia alta.
Inolte, confrontando statisticamente i dati ottenuti, possiamo notare come, su dataset molto piccoli, spesso l'aumento di tempo richiesto per aggiungere operazioni di pre, in e post-processing sia abbastanza ininfluente, ma al crescere delle dimensioni del dataset, come nel caso del Home Credit Default Risk Dataset, le operazioni arrivano a produrre anche più di **1Wh** in più rispetto alle operazioni standard. Espandendo questo concetto a dataset sempre più grandi è chiaro come l'aumento di tempo e quindi risorse necessarie a compiere azioni di pre-in-post-processing sia particolarmente verticale.
Le strategie più pesanti sembrano essere legate alle fasi di pre-in-processing, in quanto legate ad una necessaria fase di training rispetto alle operazioni di post-processing, che operano direttamente sulle predizioni fornite dai diversi modelli, andando a dimezzare, nei casi di datasets e modelli computazionalmente pesanti, i tempi e le risorse richieste.

Per concludere e fornire un'ottica ancora più dettagliata sui consumi, nella documentazione di CodeCarbon, disponibile in piè di pagina[1], si possono trovare alcuni esempi di modelli realistici, con un numero di epoche maggiore e con gestione degli iper-parametri, che raggiungono anche la produzione di 3.1KWh fino a modelli computazionalmente molto grandi che richiedono anche giorni di esecuzione per raggiungere un modello finale, consumando anche più di 30KWh. Forniamo ora una risposta alla seguente RQ:

👉 **Answer to RQ_2.** Queste informazioni ci permettono di evidenziare, ancora una volta, i consumi dei modelli di AI, e giustificando il motivo per cui, attualmente, sia uno dei maggiori argomenti di dibattito, sia in letteratura che in contesti reali, del mondo AI. Dai risultati si può notare un aumento piuttosto verticale dei consumi all'aumentare della complessità dei datasets in termini di grandezza e di features. In particolare, **AIFairness360** è la libreria, fra le due utilizzate nello studio, che sembra richiedere più risorse in termini di operazioni di pre-processing ed in-processing, rispetto alle soluzioni proposte da **Fairlearn**. L'utilizzo delle risorse e del tempo di computazione richiesto sembra essere legato alla complessità dei modelli e dei dataset utilizzati, rendendo particolarmente importante l'attenzione alla sostenibilità economica e ambientale importante quando si operano con quest'ultimi.

[1] `https://mlco2.github.io/codecarbon/model_examples.html`

4.3 RQ3: Che impatto hanno gli strumenti di Fairness sulla sostenibilità economica del prodotto?

Questa RQ nasce dalla volontà di esaminare a livello qualitativo i *trade-offs* sostenuti per ottenere dei modelli più "fair". Per poter fornire una risposta concreta a questa RQ è necessario esaminare il livello qualitativo ottenuto dai modelli *Standard* ed i modelli *Fair*, esaminando tutte le possibili complicazioni ottenute nel costruire un modello più equo.
Verranno di seguito presentati i risultati ottenuti per ogni dataset in termini di *Accuracy*, *Precision*, *Recall* ed *F1-Score* dai modelli *Standard* e dai modelli *Fair*, ottenuti sia dalla libreria **Fairlearn** che dalla libreria **AIFairness360**.
Partendo dai modelli base, addestrati sui datasets senza alcuna operazione di Fairness, nelle **Tabelle 4.22, 4.23, 4.24, 4.25** è possibile visionare i risultati ottenuti in termini di metriche di qualità dei diversi modelli nei diversi datasets.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	85%	68%	75%	62%
Random Forest	84%	66%	72%	62%
SVM	85%	67%	76%	60%
XGBoost	**87%**	**72%**	**78%**	**66%**

Tabella 4.22: Le metriche ottenute dai modelli standard sull'Adult Dataset.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	63%	69%	**93%**	55%
Random Forest	**82%**	**89%**	81%	**98%**
SVM	63%	69%	**93%**	55%
XGBoost	78%	86%	83%	89%

Tabella 4.23: Le metriche ottenute dai modelli standard sul German Credit Dataset.

Infine, dai modelli di DL realizzati sull'UTKFace dataset, si ottengono i risultati presentati in **Tabella 4.26**

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	82%	83%	84%	**82%**
Random Forest	78%	79%	83%	76%
SVM	**85%**	**86%**	**90%**	**82%**
XGBoost	70%	73%	73%	73%

Tabella 4.24: Le metriche ottenute dai modelli standard sul Heart Disease Dataset.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	65%	64%	65%	64%
Random Forest	**86%**	**87%**	**85%**	**88%**
SVM	65%	65%	65%	64%
XGBoost	73%	74%	72%	76%

Tabella 4.25: Le metriche ottenute dai modelli standard sul Home Credit Default Risk Dataset.

Modello	Accuracy	Precision	Recall	F1-Score
Resnet50	**98**%	**80**%	**98**%	**80%**
MobileNetV2	74%	74%	74%	73%

Tabella 4.26: Le metriche di qualità ottenute dai modelli di DL sull'UTKFace Dataset.

Dai risultati appena presentati è possibile effettuare delle riflessioni su quale modello si sia comportato "meglio" per ogni dataset dello studio.
In generale, possiamo notare come tutti gli algoritmi scelti, si comportino discretamente bene su tutti i dataset proposti nello studio, indice che permette di evidenziare come gli algoritmi scelti sposino il problema di classificazione posto in essere dai diversi dataset.
In particolare, analizzando le medie per ogni metrica ottenuta dai vari modelli, nonostante i risultati siano pressochè simili, possiamo ritenere il modello basato su XGBoost come il miglior modello per il problema di classificazione posto dall'Adult Dataset; per il German Credit dataset la scelta ricade sul modello basato su Random Forest che ottiene in generale i risultati migliori per ogni metrica; Per il dataset Heart Disease, torniamo su valori simili, scegliendo come modello migliore il modello basato su algoritmo SVM; infine, per il Dataset Home Credit Default Risk, ancora una volta il modello basato su Random Forest risulta essere il migliore, con risultati ben sopra la media generale dei modelli.
Tra i due modelli di DL realizzati la situazione è molto più evidente, in quanto, il modello basato su rete neurale Resnet50 ottiene valori migliori in ogni metrica, evidenziando una qualità maggiore nel problema di classificazione posto in essere dal dataset rispetto al modello basato su rete neurale MobileNetV2.

Esaminiamo ora i risultati ottenuti dai modelli realizzati con operazioni di pre-in-post processing, sfruttando le soluzioni offerte dalle librerie di Fairness, in particolare, andando ad evidenziare i casi in cui la metrica è migliorata o peggiorata. Questo ci permette di evidenziare ad occhio eventuali cambiamenti, in meglio o in peggio, delle metriche qualitative a seguito di operazioni di Fairness.
Partendo dai modelli realizzati in pre-processing, nelle **Tabelle 4.27, 4.28, 4.29, 4.30 e 4.31**, vengono presentati i risultati ottenuti dai modelli realizzati sfruttando le soluzioni offerte dalle due librerie di Fairness, anche qui vengono sfrutati gli acronimi **AIF** (AIFairness360) e **FLN** (Fairlearn) per indicare le due librerie e **LR** (Logistic Regression), **RF** (Random Forest), **SVM** (Support Vector Machine) e **XGB** (XGBoost) per indicare i modelli realizzati.

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	85%	65%(3%)↓	75%	58%(4%)↓
RF	AIF	84%	66%	71%(1%)↓	62%
SVM	AIF	84%	64%(3%)↓	76%	56%(4%)↓
XGB	AIF	**86%**(1%)↓	**70%**(2%)↓	**78%**	**63%**(3%)↓
Oggetto utilizzato: **Reweighing**					
LR	FLN	85%	68%	75%	62%
RF	FLN	84%	67%(1%)↑	71%(1%)↓	62%
SVM	FLN	85%	67%	**76%**	60%
XGB	FLN	**86%**(1%)↓	**70%**(2%)↓	**76%**(2%)↓	**65%**(1%)↓
Oggetto utilizzato: **CorrelationRemover**					

Tabella 4.27: Le metriche ottenute dai modelli con pre-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	63%	69%	90%(3%)↓	56%(1%)↑
RF	AIF	**82%**	**89%**	82%(1%)↑	**97%**(1%)↓
SVM	AIF	63%	70%(1%)↑	**91%**(2%)↓	56%(1%)↑
XGB	AIF	80%(2%)↑	87%(1%)↑	83%	91%(2%)↑
Oggetto utilizzato: **Reweighing**					
LR	FLN	63%	69%	**93%**	55%
RF	FLN	**81%**(1%)↓	**88%**(1%)↓	81%	**97%**(1%)↓
SVM	FLN	63%	69%	**93%**	55%
XGB	FLN	76%(2%)↓	84%(2%)↓	82%(1%)↓	87%(2%)↓
Oggetto utilizzato: **CorrelationRemover**					

Tabella 4.28: Le metriche ottenute dai modelli con pre-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	77%(5%)↓	77%(6%)↓	85%(1%)↑	70%(12%)↓
RF	AIF	**83%(5%)↑**	**84%(5%)↑**	**90%(7%)↑**	**79%(3%)↑**
SVM	AIF	77%(8%)↓	77%(9%)↓	85%(5%)↓	70%(10%)↓
XGB	AIF	72%(2%)↑	73%	77%(4%)↑	70%(3%)↓
Oggetto utilizzato: **Reweighing**					
LR	FLN	**85%(3%)↑**	**86%(3%)↑**	**90%(6%)↑**	**82%**
RF	FLN	77%(1%)↓	78%(q%)↓	81%(2%)↓	76%
SVM	FLN	**85%**	**86%**	**90%**	**82%**
XGB	FLN	72%(2%)↑	74%(1%)↑	75%(2%)↑	73%
Oggetto utilizzato: **CorrelationRemover**					

Tabella 4.29: Le metriche ottenute dai modelli con pre-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	65%	64%	66%(1%)↑	62%(2%)↓
RF	AIF	**87%1(%)↑**	**87%**	**86%(1%)↑**	**89%(1%)↑**
SVM	AIF	65%	64%(1%)↓	66%(1%)↑	62%(2%)↓
XGB	AIF	74%(1%)↑	74%	73%(1%)↑	75%(1%)↓
Oggetto utilizzato: **Reweighing**					
LR	FLN	65%	64%	65%	64%
RF	FLN	**86%**	**86%(1%)↓**	**85%**	**88%**
SVM	FLN	65%	65%	65%	64%
XGB	FLN	73%	74%	72%	75%(1%)↓
Oggetto utilizzato: **CorrelationRemover**					

Tabella 4.30: Le metriche ottenute dai modelli con pre-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn

Modello	Libreria	Accuracy	Precision	Recall	F1-Score
Resnet50	AIF	77%(13%)↓	77%(3%)↓	77%(21%)↓	77%(3%)↓
MobileNetV2	AIF	68%(6%)↓	64%(10%)↓	68%(6%)↓	68%(5%)↓
Oggetto utilizzato: **Reweighing**					
Resnet50	FLN	**98%**	77%(3%)↓	**98%**	77%(3%)↓
MobileNetV2	FLN	72%(2%)↓	72%(2%)↓	72%(2%)↓	71%(2%)↓
Oggetto utilizzato **CorrelationRemover**					

Tabella 4.31: Le metriche di qualità ottenute dai modelli di DL con pre-processing sull'UTK-Face Dataset. (AIF = AIFairness360, FLN = Fairlearn)

Nella fase di pre-processing è possibile evidenziare cali di qualità dell'1-6% dei modelli in tutte le metriche con alcune eccezioni, in cui i valori delle metriche invece tendono ad aumentare, ad esmepio nel Heart Disease Dataset, addirittura, il modello RF arriva ad ottenere i risultati migliori rispetto al modello SVM che nella sua implementazione standard viene preso come miglior modello, ma non solo, si può notare come la strategia offerta da AIF360 favorisca il modello RF rispetto alla controparte di Fairlearn che, invece, favorisce il modello LR portandolo ad essere il migliore fra i 4 realizzati.
Considerati i valori forniti dalle operazioni di Fairness, in relazione al calo di qualità subito dai modelli di ML, trattandosi di valori vicino all'1/2% in media, che raggiungono massimo il -5% in rari casi, o come in alcuni casi particolari nel Heart Disease Dataset dove addirittura troviamo un aumento generale delle metriche per alcuni modelli, possiamo considerare il *Trade-off* qualitativo insignificante rispetto ai valori di Fairness raggiunti a seguito del pre-processing sui dataset.
Per i modelli di DL la situazione è ben diversa, infatti, il modello *Resnet50* che otteneva grandi risultati sul dataset standard, senza operazioni di pre-processing, porta ad un livello discreto delle metriche di qualità, mentre l'altro modello basato su *MobileNetV2* segue il trend di peggioramento delle metriche, e, al netto dei risultati discreti di Fairness ottenuti, non è una strategia molto valida, su questo specifico dataset.

Per quanto riguarda la fase di in-processing, invece, data l'assenza di un tool di mitigazione di tipo in-processing della libreria Fairlearn, verranno confrontati i risultati dei soli modelli con in-processing realizzati tramite la libreria AIF360.
Nelle **Tabelle 4.32, 4.33, 4.34 e 4.35, 4.36**, vengono presentati i risultati della fase di in-processing.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	**56%**(29%)↓	**46%**(22%)↓	**33%**(42%)↓	**74%**(12%)↑
Random Forest	55%(29%)↓	**46%**(20%)↓	**33%**(39%)↓	**74%**(12%)↑
SVM	55%(30%)↓	45%(22%)↓	**33%**(43%)↓	73%(13%)↑
XGBoost	55%(32%)↓	45%(22%)↓	**33%**(45%)↓	73%(7%)↑
Oggetto utilizzato: **MetaFairClassifier**				

Tabella 4.32: Le metriche ottenute dai modelli con in-processing sull'Adult Dataset.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	78%(15%)↑	86%(17%)↑	**81%**(12%)↓	93%(38%)↑
Random Forest	**79%**(3%)↓	**88%**(1%)↓	80%(1%)↓	**97%**(1%)↓
SVM	78%(15%)↑	87%(18%)↑	**81%**(12%)↓	93%(38%)↑
XGBoost	**79%**(1%)↑	87%(1%)↑	80%(3%)↓	95%(6%)↑
Oggetto utilizzato: **MetaFairClassifier**				

Tabella 4.33: Le metriche ottenute dai modelli con in-processing sul German Credit Dataset.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	83%(1%)↑	84%(1%)↑	87%(2%)↑	**82%**
Random Forest	**85%**(7%)↑	**85%**(6%)↑	**93%**(10%)	79%(3%) ↑
SVM	83%(2%)↓	84%(2%)↓	87%(3%)↓	**82%**
XGBoost	83%(13%)↑	84%(11%)↑	90%(17%)↑	79%(6%)↑
Oggetto utilizzato: **MetaFairClassifier**				

Tabella 4.34: Le metriche ottenute dai modelli con in-processing sul Heart Disease Dataset.

Modello	Accuracy	F1-Score	Precision	Recall
Logistic Regression	65%	64%	65%	64%
Random Forest	**87%**(1%)↑	**87%**	**85%**	**88%**
SVM	65%	65%	65%	64%
XGBoost	73%	74%	72%	76%
Oggetto utilizzato: **MetaFairClassifier**				

Tabella 4.35: Le metriche ottenute dai modelli con in-processing sul Home Credit Default Risk Dataset.

Modello	Accuracy	Precision	Recall	F1-Score
Resnet50	70%(28%)↓	**63%**(17%)↓	67%(31%)↓	80%
MobileNetV2	**72%**(2%)↓	60%(14%)↓	**72%**(2%)↓	**98%**(25%)↑
Oggetto utilizzato: **MetaFairClassifier**				

Tabella 4.36: Le metriche di qualità ottenute dai modelli di DL con in-processing sull'UTKFace Dataset.

Come è possibile vedere, in particolare i modelli realizzati sull'Adult dataset con fase di in-processing, ottengono risultati davvero scarsi, cadendo anche al di sotto del 50% in alcune metriche, rendendo quasi inutilizzabili i modelli per predire nuovi dati, sicuramente la scelta di in-processing sull'Adult Dataset non è una valida possibilità. I modelli sui restanti dataset non sembrano soffrire particolarmente, seguendo anche meno il trend del pre-processing, peggiorando le proprie performance qualitative solo in rari casi e, in casi più fortuiti, addirittura le performance dei modelli migliorano. Con i valori ottenuti anche in fase di analisi di Fairness, ottenuta con questa strategia in-processing applicata, possiamo concludere che il *trade-off* qualitativo, in caso di in-processing, in alcun casi particolari, possa realmente migliorare le performance qualitative dei modelli o intaccare di poco le performance ottenute dai modelli standard, rispetto le strategie di pre-post-processing, però, in alcuni casi particolari, le performance sono peggiorate di molto, rendendo questo tipo di soluzione abbastanza altalenante.

Bisognerebbe sfruttare ulteriori soluzioni di in-processing per poter definire anche questo tipo di strategia valido, oggetto sicuramente di un futuro studio, visto l'aumento di performance notevole da parte dei modelli su alcuni datasets.
Con questo tipo di strategia, i modelli di DL non sembrano perdere troppo in termini di qualità, il trend di peggioramento delle performance di qualità rimane, ma porta i modelli in situazioni più accettabili rispetto ai modelli realizzati in pre-post-processing. Qui il *trade-off* qualitativo potrebbe anche essere accettato, con particolare attenzione ai risultati prodotti, per modelli DL che non richiedono particolari esigenze di qualità.
Infine, per la fase di post-processing, in **Tabelle 4.37, 4.41, 4.38, 4.39 e 4.40**, vengono presentati i risultati ottenuti, per maggior chiarezza e leggibilità dei risultati vengono riutilizzati gli acronimi **AIF** (AIFairness360) e **FLN** (Fairlearn) per le due librerie di Fairness e **LR** (Logistic Regression), **RF** (Random Forest), **SVM** (Support Vector Machine) e **XGB** (XGBoost) per indicare i modelli.

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	83%(2%)↓	62%(6%)↓	71%(4%)↓	55%(7%)↓
RF	AIF	83%(1%)↓	63%(3%)↓	71%(1%)↓	57%(5%)↓
SVM	AIF	83%(2%)↓	63%(4%)↓	72%(4%)↓	55%(5%)↓
XGB	AIF	**86%(1%)↓**	**69%(3%)↓**	**78%**	**62%(4%)↓**
Oggetto utilizzato: **EqOddsPostprocessing**					
LR	FLN	82%(3%)↓	55%(13%)↓	72%(3%)↓	45%(17%)↓
RF	FLN	78%(6%)↓	49%(17%)↓	60%(16%)↓	41%(21%)↓
SVM	FLN	81%(4%)↓	55%(12%)↓	71%(5%)↓	45%(15%)↓
XGB	FLN	**83%(4%)↓**	**59%(27%)↓**	**73%(5%)↓**	**49%(17%)↓**
Oggetto utilizzato: **ThresholdOptimizer**					

Tabella 4.37: Le metriche ottenute dai modelli con post-processing sull'Adult Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	62%(1%)↓	69%	91%(2%)↓	56%(1%)↑
RF	AIF	**81%(3%)↓**	**89%**	81%	**99%(1%)↑**
SVM	AIF	64%(1%)↑	70%(1%)↑	**92%(1%)↓**	57%(2%)↑
XGB	AIF	78%	86%	83%	89%
Oggetto utilizzato: **EqOddsPostprocessing**					
LR	FLN	78%(15%)↑	86%(17%)↑	**83%(10%)↓**	89%(34%)↑
RF	FLN	**79%(3%)↓**	**87%(2%)↓**	81%	**94%(4%)↓**
SVM	FLN	76%(13%)↑	85%(16%)↑	82%(11%)↓	87%(32%)↑
XGB	FLN	75%(3%)↓	84%(2%)↓	82%(1%)↓	86%(3%)↓
Oggetto utilizzato: **ThresholdOptimizer**					

Tabella 4.38: Le metriche ottenute dai modelli con post-processing sul German Credit Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	82%	83%	84%	**82%**
RF	AIF	78%	79%	83%	76%
SVM	AIF	**85%**	**86%**	**90%**	**82%**
XGB	AIF	70%	73%	73%	73%
Oggetto utilizzato: **EqOddsPostprocessing**					
LR	FLN	**65%(17%)↓**	**68%(15%)↓**	69%(15%)↓	**67%(15%)↓**
RF	FLN	63%(15%)↓	59%(20%)↓	**76%(7%)↓**	49%(27%)↓
SVM	FLN	63%(22%)↓	63%(23%)↓	70%(20%)↓	58%(24%)↓
XGB	FLN	58%(12%)↓	55%(18%)↓	68%(5%)↓	46%(27%)↓
Oggetto utilizzato: **ThresholdOptimizer**					

Tabella 4.39: Le metriche ottenute dai modelli con post-processing sul Heart Disease Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Libreria	Accuracy	F1-Score	Precision	Recall
LR	AIF	63%(2%)↓	61%(3%)↓	63%(2%)↓	60%(4%)↓
RF	AIF	**85%(1%)↓**	**85%(2%)↓**	**85%**	**85%(3%)↓**
SVM	AIF	63%(2%)↓	61%(4%)↓	64%(1%)↓	59%(5%)↓
XGB	AIF	72%(1%)↓	72%(2%)↓	71%(1%)↓	72%(4%)↓
Oggetto utilizzato: **EqOddsPostprocessing**					
LR	FLN	63%(2%)↓	64%	62%(3%)↓	66%(2%)↑
RF	FLN	**74%(12%)↓**	**77%(10%)↓**	70%(15%)↓	**87%(1%)↓**
SVM	FLN	64%(1%)↓	64%(1%)↓	63%(2%)↓	65%(1%)↑
XGB	FLN	72%(1%)↓	72%(2%)↓	**72%**	73%(3%)↓
Oggetto utilizzato: **ThresholdOptimizer**					

Tabella 4.40: Le metriche ottenute dai modelli con post-processing sul Home Credit Default Risk Dataset. (LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Accuracy	Precision	Recall	F1-Score
Resnet50	48%(50%)↓	**60%(20%)↓**	47%(51%)↓	**84%(4%)↓**
MobileNetV2	**52%(22%)↓**	42%(32%)↓	**49%(25%)↓**	37%(36%)↓
Oggetto utilizzato: **EqOddsPostprocessing**				

Tabella 4.41: Le metriche di qualità ottenute dai modelli di DL con post-processing sull'UTK-Face Dataset. (AIF = AIFairness360, FLN = Fairlearn)

Anche in fase di post-processing, il trend rimane lo stesso, c'è un fenomeno generalizzato di peggioramento delle metriche. In particolare, si può notare come la soluzione di post-processing offerta dalla libreria Fairlearn peggiori vertiginosamente le metriche ottenute sul Heart Disease Model, mentre la controparte di AIF360 sullo stesso dataset non subisca alcuna modifica.

Possiamo concludere, dati i risultati di Fairness ottenuti da questo tipo di strategia, evidenziati nella RQ1, che il *trade-off* qualitativo ottenuto in fase di post-processing, per le soluzioni previste dalle due librerie, sia insignificante, favorendo un comportamento molto più equo dei modelli ad un minimo costo di qualità ottenuta nella maggioranza dei casi.

Per i modelli i DL, come nel caso del pre-processing, in maniera anche più eclatante, le performance qualitative dei modelli calano drasticamente, raggiungendo valori non più utili per poter considerare il modello come soluzione utile. Anche qui, il *trade-off* qualitativo rispetto ai risultati di Fairness ottenuti, non consente di poter utilizzare queste strategie per realizzare modelli DL di qualità equi.
Analizziamo, inoltre, i tempi necessari richiesti per eseguire le fasi di pre, in e post-processing e di training e testing dei modelli, per poter confrontare il possibile overhead computazionale che le strategie offerte dalle due librerie hanno rispetto alle fasi di training e testing di modelli standard.
Andremo ora ad esaminare i tempi richiesti per ogni tipo di operazione di pre, in e post-processing, nelle **Tabelle 4.42, 4.43 e 4.44**, infatti, troviamo rispettivamente i risultati di AIF360 e Fairlearn in fase di pre-processing, in-processing e post-processing per ognuna delle 10 iterazioni, e, nelle **Tabelle 4.45 e 4.46** i risultati dei modelli DL, andando ad evidenziare miglioramenti o peggioramenti nel tempo di esecuzione richiesto.

Dataset	Tempo medio	Tempo medio
Adult	4.94	5.64(14.17%)↑
German	1.71	1.95(14.04%)↑
Heart	1.53	1.62(5.88%)↑
Home	35.05	40.2(14.70%)↑
Oggetto utilizzato: **Reweighing**		
Adult	4.94	6.72(36.03%)↑
German	1.71	2.35(37.43%)↑
Heart	1.53	1.99(30.07%)↑
Home	35.05	43.65(24.54%)↑
Oggetto utilizzato: **CorrelationRemover**		

Tabella 4.42: I tempi medi necessari ad eseguire i vari script con pre-processing (AIF360-Fairlearn) sui diversi datasets (in secondi).

Dataset	Tempo medio	Tempo medio
Adult	4.94	10.59(114.37%)↑
German	1.71	6.48(278.95%)↑
Heart	1.53	1.62(5.88%)↑
Home	35.05	39.65(13.12%)↑
Oggetto utilizzato: **MetaFairClassifier**		

Tabella 4.43: I tempi medi necessari ad eseguire i vari script con in-processing (AIF360) sui diversi datasets (in secondi).

Dataset	Tempo medio	Tempo medio
Adult	4.94	2.16(56.28%)↓
German	1.71	1.61(5.85%)↓
Heart	1.53	1.56(1.96%)↑
Home	35.05	3.35(90.44%)↓
Oggetto utilizzato: **EqOddsPostprocessing**		
Adult	4.94	6.25(26.52%)↑
German	1.71	3.23(88.89%)↑
Heart	1.53	2.38(55.56%)↑
Home	35.05	43.7(24.68%)↑
Oggetto utilizzato: **ThresholdOptimizer**		

Tabella 4.44: I tempi necessari ad eseguire i vari script con post-processing (AIF360-Fairlearn) sui diversi datasets (in secondi).

Modello	Tempo medio	Tempo medio
Resnet50	45.65	46.58(2.04%)↑
MobileNetV2	44.95	46.12(2.60%)↑
Oggetto utilizzato: **Reweighing**		
Resnet50	45.65	46.58(2.04%)↑
MobileNetV2	44.95	46.12(2.60%)↑
Oggetto utilizzato: **CorrelationRemover**		

Tabella 4.45: I tempi necessari ad eseguire i vari script dei modelli DL con pre-processing (AIF360-Fairlearn) sull'UTKFace Dataset (in secondi).

Modello	Tempo medio	Tempo medio
Resnet50	45.65	45.94(0.64%)↑
MobileNetV2	44.95	46.22(2.83%)↑
Oggetto utilizzato: **MetaFairClassifier**		
Resnet50	45.65	14.8(67.58%)↓
MobileNetV2	44.95	11.8(73.75%)↓
Oggetto utilizzato: **EqOddsPostprocessing**		

Tabella 4.46: I tempi necessari ad eseguire i vari script dei modelli DL con in-processing e post-processing (AIF360) sull'UTKFace Dataset (in secondi).

Com'è possibile vedere, ovviamente, utilizzare delle soluzioni che presentano pre, in e post-processing rispetto ad utilizzare i modelli e i datasets *as-is* comporta maggior overhead computazionale, questo tipo di risultato però non sembra particolarmente influente sui tempi, si parla di circa 1-2s di aumento massimo nei dataset più piccoli, mentre in datasets molto grandi arriviamo anche a 10s massimi di ritardo, sono comunque dei valori accettabili se confrontiamo i risultati in termini di Fairness ottenuti nella RQ1. Possiamo, inoltre, evidenziare come il post-processing di AIF360 riesca a migliorare sui dataset grandi come l'Home Credit Default Risk, in quanto, trattandosi di una operazione che modifica le sole predizioni fornite dai modelli, non richiede alcuna fase di training e testing, riducendo di molto il tempo richiesto; la controparte di Fairlearn, invece, essendo un vero e proprio oggetto che effettua una fase di re-training del modello per fornire delle predizioni maggiormente eque, richiede più tempo rispetto le soluzioni standard proposte.
Valutiamo, infine, i pesi dei modelli ottenuti dalle varie esecuzioni.
In **Tabella 4.47, 4.48, 4.49, 4.50 e 4.51** sono presentati i risultati.

Modello	Lib.	Peso (MB)	Peso (MB)	Peso (MB)	Peso (MB)
LR	AIF	0.0069	0.0069	0.0069	-
RF	AIF	69.1	72.3(4.63%)↑	14.0(79.74%)↓	-
SVM	AIF	0.0068	0.0068	0.0067(1.47%)↓	-
XGB	AIF	0.27	0.27	0.21(22.22%)↓	-
LR	FLN	0.0069	0.0069	-	0.19(2654%)↑
RF	FLN	69.1	69.1	-	69.2(0.15%)↑
SVM	FLN	0.0068	0.0068	-	0.19(2694%)↑
XGB	FLN	0.27	0.27	-	0.44(62.96%)↑

Tabella 4.47: I pesi (in MB) ottenuti dai vari modelli sull'Adult Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Lib.	Peso (MB)	Peso (MB)	Peso (MB)	Peso (MB)
LR	AIF	0.0044	0.0044	0.0045(2.27%)↑	-
RF	AIF	3.3	3.3	1.8(45.46%)↓	-
SVM	AIF	0.0043	0.0043	0.0043	-
XGB	AIF	0.20	0.20	0.15(25%)↓	-
LR	FLN	0.0044	0.0044	-	0.35(7855%)↑
RF	FLN	3.3	3.3	-	3.4(3.03%)↑
SVM	FLN	0.0043	0.0043	-	0.35(8040%)↑
XGB	FLN	0.20	0.20	-	0.55(175%)↑

Tabella 4.48: I pesi (in MB) ottenuti dai vari modelli sul German Credit Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Lib.	Peso (MB)	Peso (MB)	Peso (MB)	Peso (MB)
LR	AIF	0.0016	0.0016	0.0017(6.25%)↑	-
RF	AIF	0.75	0.75	0.45(40%)↓	-
SVM	AIF	0.0015	0.0015	0.0015	-
XGB	AIF	0.12	0.12	0.09(25%)↓	-
LR	FLN	0.0016	0.0016	-	0.17(10525%)↑
RF	FLN	0.75	0.75	-	0.88(17.33%)↑
SVM	FLN	0.0015	0.0015	-	0.17(11233%)↑
XGB	FLN	0.12	0.12	-	0.29(141.67%)↑

Tabella 4.49: I pesi (in MB) ottenuti dai vari modelli sul Heart Disease Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Lib.	Peso (MB)	Peso (MB)	Peso (MB)	Peso (MB)
LR	AIF	0.0043	0.0043	0.0043	-
RF	AIF	204.5	213.4(4.35%)↑	204.7(0.01%)↑	-
SVM	AIF	0.0041	0.0041	0.0042(2.44%)↑	-
XGB	AIF	0.43	0.43	0.43	-
LR	FLN	0.0043	0.0043	-	0.18(4086%)↑
RF	FLN	204.5	204.5	-	204.7(0.01%)↑
SVM	FLN	0.0041	0.0041	-	0.18(4290%)↑
XGB	FLN	0.43	0.43	-	0.61(41.86%)↑

Tabella 4.50: I pesi (in MB) ottenuti dai vari modelli sul Home Credit Default Risk Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing.(LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, XGB = XGBoost, AIF = AIFairness360, FLN = Fairlearn)

Modello	Lib.	Peso (MB)	Peso (MB)	Peso (MB)	Peso (MB)
RN50	AIF	90.4	90.4	90.4	90.4
MNV2	AIF	9.1	9.1	9.1	9.1
RN50	FLN	90.4	90.4	-	90.4
MNV2	FLN	9.1	9.1	-	9.1

Tabella 4.51: I pesi (in MB) ottenuti dai vari modelli sull'UTKFace Dataset (AIF360-Fairlearn) in ordine di standard-pre-in-post-processing. (AIF = AIFairness360, FLN = Fairlearn, RN50 = ResNet50, MNV2 = MobileNetV2)

Com'è possibile osservare dai pesi ottenuti, i modelli basati sull'algoritmo **RandomForest** raggiungono pesi relativamente grandi rispetto ai valori ottenuti dagli altri modelli. In particolare i modelli basati su **SVM e MobileNetV2** sembrano essere di gran lunga i più leggeri per ML e DL. Si può notare un aumento dei pesi solo in corrispondenza dei modelli più grandi, a quanto pare, l'aumento sembra essere influente solo all'aumentare del peso, fino ad un valore abbastanza grande come per le soluzioni di RandomForest, che superano anche i *200MB*, segno che le operazioni abbiano impatto anche sul peso dei modelli realizzati, ma, date le dimensioni molto piccole della maggioranza dei modelli, non si possono trarre ulteriori conclusioni.
Per concludere, viene presentata una risposta alla RQ:

👉 **Answer to RQ$_3$.** Con i risultati di Fairness ottenuti, le fasi di pre-processing e di post-processing producano modelli più equi con minima riduzione di qualità del modello, il che può portare a realizzare con maggiore attenzione anche modelli di ML più equi, inserendo, da subito, nel contesto di creazione di un modello di ML anche la componente di equità del modello, dei dati e delle predizioni stesse. In particolare, bisognerebbe evolvere le strategie di in-processing fornite, per fare si che il livello di qualità dei modelli non risenta troppo rispetto ai risultati di Fairness ottenuti, che rendono i modelli equi realizzati con in-processing meno appetibili delle controparti con pre e post-processing. In relazione ai pesi, i modelli ottenuti dall'algoritmo **RandomForest** sono sistematicamente più grandi, rendendo l'algoritmo poco consigliato al fine di ottenere modelli equi anche dal punto di vista di risorse richieste. Mentre, per i modelli di DL, le strategie utilizzate e le librerie non sembrano offrire dei veri e propri risultati qualitativamente validi ai fini di realizzare un modello di qualità che presenta anche maggior equità nel comportamento, favorendo, in termini di pesi, i modelli basati su rete neurale **MobileNetV2**. Questo problema potrebbe essere legato solo problema di classificazione presentato, in quanto, parlando di image recognition, parliamo di dati di input in formato diverso da quello testuale, e questa differenza di input potrebbe rappresentare un motivo per cui le due librerie e le varie soluzioni proposte non abbiano funzionato correttamente. Ulteriori studi su più datasets e tipo di problema di classificazione sono necessari per raggiungere una conclusione soddisfacente in campo di DL.

CAPITOLO 5

Conclusioni

Per concludere questo studio, riflettiamo sui risultati e le risposte alle varie RQ proposte. Nello studio, è stato evidenziato come, sfruttando diverse librerie di Fairness, i risultati possano portare sostanzialmente ad ottenere dei modelli differenti con diversi risultati di equità. In particolare, alcune delle strategie proposte sono state in grado di ottenere dei risultati generalizzabili, con pochi e rari casi in cui le i valori di Fairness ottenuti a seguito di una specifica operazione di mitigazione sul dataset o modello abbia ottenuto risultati peggiori. Sono rari i casi in cui le soluzioni proposte abbiano realmente fallito o agito al di sotto delle aspettative, questi casi isolati non sono stati studiati ed approfonditi nello studio, inoltre, non è stato possibile sfruttare entrambe le librerie proposte in ambito di DL, in particolare in ambito di Image Recognition, in cui i dati presentati ai modelli non sono rappresentati in forma testuale tradizionale quanto come immagini, questo specifico modo di rappresentare i dati, unito alla diversa implementazione dei modelli sfruttando artefatti software diversi come le reti neurali, ha portato ad ottenere in ambito di Fairness, risultati al di sotto delle aspettative e del comportamento generale che le stesse soluzioni hanno ottenuto su modelli con input testuale.

Queste ultime dinamiche e situazioni riscontrate, aprono un mondo di possibili spunti futuri, sia per gli strumenti utilizzati, in quanto al giorno d'oggi, i software, architetture e framework di Fairness aumentano a macchia d'olio e confrontare i risultati di questo studio con nuove tecniche e infrastrutture è indubbiamente il prossimo step da compiere sia con ulteriori sulle singole dinamiche rare in cui gli strumenti hanno fallito o non hanno ottenuto i risultati sperati, andando ad esaminare il perché queste soluzioni non abbiano funzionato come previsto, riproponendo lo studio su diversi input e datasets più simili ai casi fallimentari delle soluzioni, per verificare il possibile ripetersi del fallimento e definire concretamente cosa non ha funzionato e perché. Inoltre, lo studio stesso potrebbe essere in futuro ampliato con nuove metriche di valutazione, sia qualitativa che di equità, andando ad approfondire di più le dinamiche di correlazione fra la qualità di un modello e la sua equità nel comportamento. Per concludere, bisognerebbe esplorare molto di più l'ambito di sostenibilità, che ad oggi, in mancanza di linee guida chiare e sopratutto assenza di studi nelle ulteriori sfaccettature della sostenibilità stessa, non permette di avere un idea chiara di come un modello sia "sostenibile" se non da un punto di vista energetico, di sicuro, però, lo studio qui realizzato rappresenta una base di partenza solida per poter ampliare il discorso di Fairness dei modelli e di sostenibilità di quest'ultimi e di tutte le piccole dinamiche che possono esistere e correlare il discorso di qualità e di equità, andando ad abbattere lo stereotipo per cui per realizzare un modello ed una soluzione AI *ground-breaking* al giorno d'oggi sia implicitamente necessario realizzare modelli poco "equi" nel loro comportamento per massimizzare i risultati e, sopratutto, di richiedere necessariamente una quantità di risorse, economiche, ambientali e tecnologiche insostenibili.

Bibliografia

[1] G. Velarde, "Artificial intelligence and its impact on the fourth industrial revolution: A review," 2020. (Citato a pagina 1)

[2] A. Prahl and W. W. P. Goh, ""rogue machines" and crisis communication: When ai fails, how do companies publicly respond?" *Public Relations Review*, vol. 47, no. 4, p. 102077, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0363811121000709 (Citato alle pagine 1 e 2)

[3] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data - ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2021. (Citato a pagina 1)

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, jul 2021. [Online]. Available: https://doi.org/10.1145/3457607 (Citato alle pagine 1, 10, 11, 13 e 14)

[5] J. Gu and D. Oelke, "Understanding bias in machine learning," 2019. (Citato a pagina 1)

[6] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2018. New York, NY, USA:

Association for Computing Machinery, 2018, p. 754–759. [Online]. Available: https://doi.org/10.1145/3236024.3264838 (Citato a pagina 2)

[7] Wikipedia, "Intelligenza artificiale — wikipedia, l'enciclopedia libera," 2023, [Online; in data 5-settembre-2023]. [Online]. Available: http://it.wikipedia.org/w/index.php?title=Intelligenza_artificiale&oldid=135101210 (Citato a pagina 5)

[8] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current Reviews in Musculoskeletal Medicine*, vol. 13, no. 1, pp. 69–76, Feb 2020. [Online]. Available: https://doi.org/10.1007/s12178-020-09600-8 (Citato alle pagine 5 e 6)

[9] A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, "Artificial intelligence in medicine," *Ann R Coll Surg Engl*, vol. 86, no. 5, pp. 334–338, sep 2004. (Citato a pagina 5)

[10] M. Somalvico *et al.*, *Intelligenza artificiale*. Scienza & vita nuova, 1987. (Citato a pagina 5)

[11] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997. (Citato a pagina 5)

[12] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019. (Citato a pagina 5)

[13] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b*, vol. 4, pp. 51–62, 2017. (Citato alle pagine 6 e 7)

[14] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017. (Citato alle pagine 6 e 7)

[15] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007. (Citato a pagina 7)

[16] R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *2017 International Conference on Machine Learning and Data Science (MLDS)*, 2017, pp. 37–43. (Citato a pagina 7)

[17] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software engineering for ai-based systems: A survey," *ACM Trans. Softw. Eng. Methodol.*, vol. 31, no. 2, apr 2022. [Online]. Available: https://doi.org/10.1145/3487043 (Citato alle pagine 7 e 8)

[18] E. Nascimento, A. Nguyen-Duc, I. Sundbø, and T. Conte, "Software engineering for artificial intelligence and machine learning software: A systematic literature review," 2020. (Citato a pagina 7)

[19] J. Bhalla, S. C. Cook, and D. J. Harvey, "A conceptual framework for the se of ai-intensive systems (se4ai) – considering data through the life-cycle," *INCOSE International Symposium*, vol. 33, no. 1, pp. 1333–1356, 2023. [Online]. Available: https://incose.onlinelibrary.wiley.com/doi/abs/10.1002/iis2.13085 (Citato alle pagine 7 e 9)

[20] T. McDermott, D. DeLaurentis, P. Beling, M. Blackburn, and M. Bone, "Ai4se and se4ai: A research roadmap," *INSIGHT*, vol. 23, no. 1, pp. 8–14, 2020. [Online]. Available: https://incose.onlinelibrary.wiley.com/doi/abs/10.1002/inst.12278 (Citato a pagina 8)

[21] K. Pepe and N. Hutchison, "Ai4se and se4ai: Setting the roadmap toward human-machine co-learning," *INSIGHT*, vol. 25, no. 4, pp. 80–84, 2022. [Online]. Available: https://incose.onlinelibrary.wiley.com/doi/abs/10.1002/inst.12417 (Citato a pagina 8)

[22] S. Masuda, K. Ono, T. Yasue, and N. Hosokawa, "A survey of software quality for machine learning applications," in *2018 IEEE International Conference on*

Software Testing, Verification and Validation Workshops (ICSTW), 2018, pp. 279–284. (Citato a pagina 8)

[23] G. Fujii, K. Hamada, F. Ishikawa, S. Masuda, M. Matsuya, T. Myojin, Y. Nishi, H. Ogawa, T. Toku, S. Tokumoto, K. Tsuchiya, and Y. Ujita, "Guidelines for quality assurance of machine learning-based artificial intelligence," *International Journal of Software Engineering and Knowledge Engineering*, vol. 30, no. 11n12, pp. 1589–1606, 2020. [Online]. Available: https://doi.org/10.1142/S0218194020400227 (Citato a pagina 8)

[24] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 2019, pp. 291–300. (Citato a pagina 8)

[25] I. Zliobaite, "A survey on measuring indirect discrimination in machine learning," 2015. (Citato alle pagine 10 e 11)

[26] R. Allen and D. Masters, "Artificial intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making," *ERA Forum*, vol. 20, no. 4, pp. 585–598, Mar 2020. [Online]. Available: https://doi.org/10.1007/s12027-019-00582-w (Citato a pagina 11)

[27] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, p. eaao5580, 2018. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.aao5580 (Citato a pagina 12)

[28] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–7. [Online]. Available: https://doi.org/10.1145/3194770.3194776 (Citato alle pagine 12 e 13)

[29] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf (Citato a pagina 12)

[30] S. Biswas and H. Rajan, "Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 642–653. [Online]. Available: https://doi.org/10.1145/3368089.3409704 (Citato alle pagine 13 e 15)

[31] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018. (Citato alle pagine 14, 15 e 27)

[32] S. McGuire, E. Schultz, B. Ayoola, and P. Ralph, "Sustainability is stratified: Toward a better theory of sustainable software engineering," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023, pp. 1996–2008. (Citato alle pagine 15 e 16)

[33] R. Chitchyan, C. Becker, S. Betz, L. Duboc, B. Penzenstadler, N. Seyff, and C. C. Venters, "Sustainability design in requirements engineering: State of practice," in *Proceedings of the 38th International Conference on Software Engineering Companion*, ser. ICSE '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 533–542. [Online]. Available: https://doi.org/10.1145/2889160.2889217 (Citato alle pagine 15 e 16)

[34] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, "Measuring the carbon intensity of ai in cloud instances," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York,

NY, USA: Association for Computing Machinery, 2022, p. 1877–1894. [Online]. Available: https://doi.org/10.1145/3531146.3533234 (Citato a pagina 16)

[35] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," 2020. (Citato a pagina 16)

[36] C. NICODEME, "Ai legitimacy for sustainability," in *2021 IEEE Conference on Technologies for Sustainability (SusTech)*, 2021, pp. 1–5. (Citato a pagina 16)

[37] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20. (Citato a pagina 21)

[38] L. E. Celis and V. Keswani, "Improved adversarial learning for fair classification," *ArXiv*, vol. abs/1901.10443, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59336159 (Citato a pagina 22)

[39] P. Awasthi, M. Kleindessner, and J. Morgenstern, "Effectiveness of equalized odds for fair classification under imperfect group information," *ArXiv*, vol. abs/1906.03284, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:182952909 (Citato a pagina 22)

[40] A. Rezaei, R. Fathony, O. Memarrast, and B. D. Ziebart, "Fairness for robust log loss classification," in *AAAI Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:189928055 (Citato a pagina 22)

[41] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77. (Citato a pagina 22)

[42] Y. Ahn and Y. Lin, "Fairsight: Visual analytics for fairness in decision making," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 1086–1095, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199064395 (Citato a pagina 22)

[43] X. Wang and H. Huang, "Approaching machine learning fairness through adversarial network," *ArXiv*, vol. abs/1909.03013, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202233546 (Citato a pagina 22)

[44] Y. Wu, L. Zhang, and X. Wu, "On discrimination discovery and removal in ranked data using causal graph," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3747383 (Citato a pagina 22)

[45] M. Anna, inversion, KirillOdintsov, and K. Martin, "Home credit default risk," 2018. [Online]. Available: https://kaggle.com/competitions/home-credit-default-risk (Citato a pagina 22)

[46] A. Janosi, W. Steinbrunn, M. Pfisterer, , and R. Detrano, "Heart Disease," UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C52P4X. (Citato a pagina 23)

[47] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/ (Citato a pagina 27)

Ringraziamenti

Ai miei genitori Pasquale ed Elisabetta e a mio fratello Matteo,
la mia famiglia, che mi ha insegnato cos'è il sacrificio, cosa significhi credere in qualcosa e scommetterci. Spero che questo mio traguardo possa essere un modo per ripagare i sacrifici e le speranze risposte.

Ai miei nonni, Angelo, Antonietta, Ciro e Vittoria,
la mia guida, il mio faro, da sempre e per sempre, mi avete insegnato la bellezza della serenità e di come vivere al meglio ogni momento, anche quando la vita fa' di tutto per bloccarci la strada.

A zia Pia e zia Sissy,
mi avete insegnato cosa significa amare le proprie passioni e continuare a crederci, anche quando non sembra esserci via di uscita. Questo traguardo è simbolo di aver creduto in me fino ad oggi con la stessa passione e insistenza.

A Grazia,
a te che non hai mai smesso di scommettere su di me, a te che hai scelto me, anche quando di me non c'era nulla, a te che senza "te" non ci sarebbe "me". Ci sono tante cose che vorrei dire che potrebbero descriverci ma che alla fine mi bastano due parole, *Ti amo*.

A Salvatore, Maria, Liliana, Andrea, Marco e Nicola,
la mia seconda famiglia, diventati ormai una parte di me, una parte su cui potrò sempre contare, e in questo traguardo c'è tanto me quanto le giornate spensierate, le giornate no e le giornate difficili passate insieme.

A Luigi, Mihail, Giuseppe, Antonio N., Ludovica, Antonio D.,
ognuno con la propria storia, ognuno con la propria vita, ognuno con i propri problemi, eppure, ne tanti giri della vita, ci siamo trovati proprio qui, nel momento giusto a condividere questo percorso con le persone giuste.

Agli amici, parenti, conoscenti, a chiunque stia leggendo e a chiunque altro incontrerò nel mio percorso, questo traguardo è un grazie anche per voi che rendete la vita uno splendido viaggio di cui non sarò mai stanco. *Grazie!*