UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Analisi di Algoritmi di Fairness Pre-processing in un Ambiente Federato Eterogeneo

RELATORE

Prof. Fabio Palomba

Dott. Gianmario Voria

Università degli Studi di Salerno

CANDIDATO

Alessandro Pinto

Matricola: 0512116318

Anno Accademico 2024-2025

Questa tesi è stata realizzata nel

It always seems impossible until it's done.

-Nelson Mandela

Abstract

L'impiego crescente di sistemi di Intelligenza Artificiale in domini critici ha reso improrogabile la necessità di garantirne l'equità (fairness). Parallelamente, l'Apprendimento Federato (*Federated Learning*, FL) si è affermato come paradigma chiave per l'addestramento di modelli nel rispetto della privacy, operando su dati decentralizzati. Tuttavia, emerge una lacuna critica nello stato dell'arte: come si comportano le tecniche di mitigazione del bias, concepite per dati centralizzati, quando vengono adattate e implementate in un ambiente federato realistico, caratterizzato da eterogeneità dei dati (Non-IID) e da stretti vincoli che impediscono la condivisione di qualsiasi statistica tra i client?

Questa tesi affronta direttamente tale problema attraverso una rigorosa analisi sperimentale. L'obiettivo è adattare, implementare e valutare l'efficacia di tre noti algoritmi di fairness pre-processing: *Reweighting*, *Optimized Preprocessing* e *Disparate Impact Remover* in un sistema federato puramente locale, studiandone il comportamento al variare del grado di eterogeneità dei dati.

I risultati ottenuti su due dataset di benchmark (*Adult* e *COMPAS*) dimostrano che non esiste una tecnica universalmente superiore. L'eterogeneità dei dati si è rivelata un fattore decisivo nel determinare l'algoritmo più performante. In questo quadro, la tecnica di *Reweighting* ha esibito la maggiore robustezza, garantendo il compromesso più stabile ed equilibrato tra accuratezza e fairness attraverso i diversi scenari.

In conclusione, questo lavoro fornisce una delle prime evidenze empiriche che quantificano il trade-off accuratezza-fairness per tecniche di pre-processing applicate in modo strettamente locale in contesti federati.

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

L'adozione di algoritmi di machine learning (ML) in ambiti decisionali critici, dalla finanza alla sanità, ha portato alla luce una sfida etica fondamentale: il **bias algoritmico**. I modelli di ML possono infatti replicare e amplificare i pregiudizi presenti nei dati di addestramento, producendo risultati sistematicamente iniqui verso determinati gruppi sociali. Parallelamente, la crescente sensibilità verso la protezione dei dati ha favorito l'ascesa dell'Apprendimento Federato (*Federated Learning*, FL), un paradigma che consente di addestrare modelli su dati decentralizzati senza che questi lascino mai il dispositivo dell'utente. Questo lavoro di tesi si colloca all'intersezione di queste due sfide cruciali: come possiamo garantire l'equità (*fairness*) dei modelli di intelligenza artificiale quando questi vengono addestrati in un ambiente federato, decentralizzato e intrinsecamente eterogeneo?

1.1 Contesto, Motivazioni e Obiettivi

Il problema della fairness nei sistemi di IA ha conseguenze concrete e profonde. Un modello di credit scoring che discrimina in base al genere può negare opportunità finanziarie, mentre un sistema di giustizia predittiva iniquo può influenzare sentenze e libertà personali. Sebbene la letteratura scientifica offra un'ampia gamma di tecni-

che per mitigare tali rischi in contesti centralizzati, l'avvento dell'Apprendimento Federato impone una profonda riconsiderazione della loro validità e applicabilità.

L'ambiente federato introduce una complessità peculiare: i dati non sono solo distribuiti per tutelare la privacy, ma anche tipicamente eterogenei (o **Non-IID**), riflettendo la diversità del mondo reale. Questa condizione genera un problema di ricerca critico e ancora poco esplorato: come si comportano le strategie di fairness, concepite per una visione globale dei dati, quando vengono applicate su "silos" di dati locali, parziali e potenzialmente distorti? L'applicazione di una correzione a livello locale può portare a un modello globale effettivamente equo, o rischia di introdurre nuove e impreviste forme di bias?

Lo stato dell'arte attuale, come verrà approfondito nel prossimo capitolo, offre poche risposte. Questa tesi si propone di colmare tale lacuna. L'obiettivo principale è **adattare, implementare e valutare in modo sistematico l'efficacia di alcuni dei più noti algoritmi di fairness pre-processing in un contesto di apprendimento federato, analizzando come la loro performance cambi al variare del grado di eterogeneità dei dati**.

1.2 Contributi e Risultati Principali

Il contributo principale di questa tesi è una rigorosa **analisi empirica** che fa luce sul comportamento delle più comuni tecniche di fairness pre-processing in scenari federati realistici, dove i client operano in modo isolato. Laddove la letteratura presenta incertezze, questo lavoro fornisce prove concrete e quantificabili. I risultati ottenuti hanno evidenziato diversi aspetti chiave:

- Esiste un trade-off fondamentale tra accuratezza e fairness, ma questo non è inviolabile. In condizioni specifiche di alta eterogeneità, la mitigazione del bias può portare a un inaspettato e significativo miglioramento dell'accuratezza.
- Non esiste un algoritmo di fairness che sia universalmente superiore. La scelta della tecnica ottimale dipende in modo critico dal dataset e, soprattutto, dal livello di eterogeneità dei dati distribuiti tra i client.

- La tecnica *Reweighting* è emersa come l'approccio più robusto e affidabile, offrendo un compromesso stabile ed efficace nella maggior parte degli scenari. Al contrario, *Disparate Impact Remover* ha mostrato un comportamento altamente volatile, legato a complesse interazioni con il modello di classificazione, risultando eccezionale in un contesto e del tutto inadatto in un altro.

Questi risultati forniscono una guida pratica per la progettazione di sistemi di IA equi e federati e pongono le basi per future ricerche nel campo.

1.3 Struttura della Tesi

Il resto della tesi è organizzato come segue:

Capitolo 2: *Stato dell'Arte.* Vengono approfonditi i concetti di Apprendimento Federato e fairness algoritmica, discutendo le principali tecniche di mitigazione e le sfide specifiche che emergono nel contesto FL.

Capitolo 3: *Metodo di Ricerca.* Viene descritto in dettaglio l'assetto sperimentale, inclusi i dataset (*Adult* e *COMPAS*), le tecniche adattate, la configurazione del sistema federato e il metodo per simulare diversi livelli di eterogeneità.

Capitolo 4: *Risultati e Discussione.* Vengono presentati e analizzati i risultati quantitativi, interpretando le performance delle tecniche alla luce delle domande di ricerca.

Capitolo 5: *Conclusioni e Sviluppi Futuri.* Il capitolo finale riassume i contributi della tesi, ne riconosce i limiti e propone direzioni promettenti per la ricerca futura.

CAPITOLO 2

Stato dell'arte

2.1 Federated Learning

2.1.1 Definizione

L'Apprendimento Federato (*Federated Learning*) [13] è un paradigma di apprendimento automatico distribuito che consente l'addestramento di un modello globale su dati decentralizzati, residenti su una moltitudine di client, senza che questi vengano condivisi. Il processo è coordinato da un server centrale che orchestra cicli di addestramento iterativi: il modello corrente viene trasmesso ai client, i quali lo perfezionano localmente sui propri dati. Successivamente, solo gli aggiornamenti dei parametri del modello vengono comunicati al server, che li aggrega per produrre una nuova versione del modello globale. Questo meccanismo garantisce la **privacy** degli utenti e facilita la conformità a normative stringenti sulla protezione dei dati, come il GDPR o l'HIPAA.

Il diagramma in Figura 2.1 illustra il ciclo iterativo del FL, che si articola in tre passaggi principali:

- **Step 1: Distribuzione del Modello.** Il server centrale distribuisce il modello globale corrente a un sottoinsieme di client.

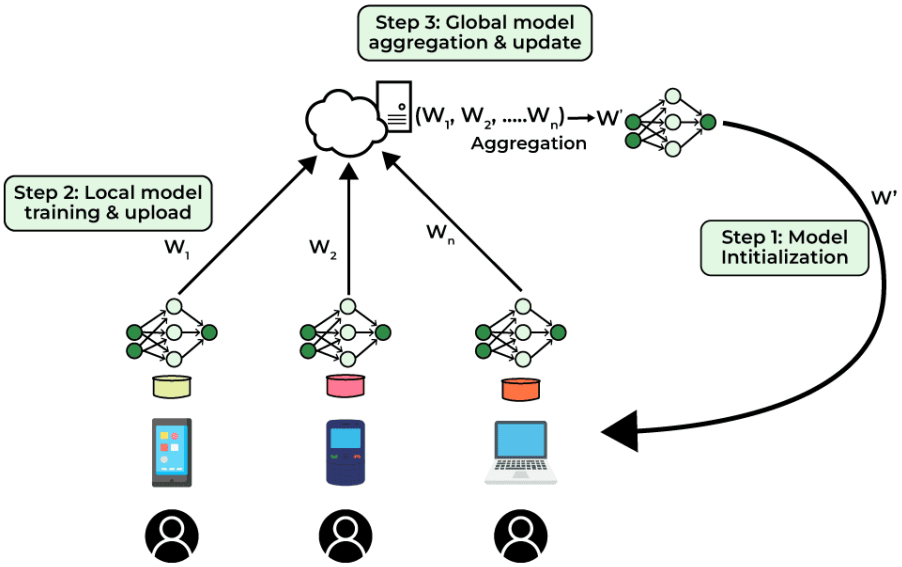

Figura 2.1: Processo iterativo del Federated Learning.

- **Step 2: Addestramento Locale.** Ciascun client addestra il modello ricevuto utilizzando il proprio dataset privato, senza mai esporre i dati.
- **Step 3: Aggregazione.** I client inviano solo gli aggiornamenti (es. i gradienti o i pesi aggiornati) al server, che li aggrega (es. con l'algoritmo FedAvg [13]) per migliorare il modello globale, pronto per il ciclo successivo.

2.1.2 Vantaggi e Sfide

I principali vantaggi del FL includono **privacy by design**, **scalabilità** e **efficienza di comunicazione**. Tuttavia, il paradigma introduce sfide uniche, la più importante delle quali è la gestione di dati **Non-IID** (non indipendenti e identicamente distribuiti). L'eterogeneità statistica tra i dati dei client può degradare le performance del modello e rallentarne la convergenza.[18]

2.2 Fairness nel Machine Learning

2.2.1 Definizione di Fairness

Nel contesto del machine learning, la **fairness** (equità) si riferisce all'assenza di discriminazioni o bias ingiustificati nelle decisioni di un modello, in relazione a uno o più **attributi protetti** (es., genere, etnia, età) [4]. Sebbene esistano molteplici definizioni, questo studio si concentra sulla **fairness di gruppo** [7], che richiede che i gruppi definiti da un attributo sensibile A ottengano risultati statisticamente comparabili. Una delle formalizzazioni più note è l'indipendenza statistica, che impone che la predizione del modello $\hat{Y}$ sia indipendente da A:

$$\hat{Y} \perp A$$

Questo concetto è alla base di diverse metriche quantitative per la valutazione della fairness.

2.2.2 Approcci alla Mitigazione del Bias

Le strategie per mitigare il bias si dividono tipicamente in tre categorie, a seconda della fase del pipeline di machine learning in cui intervengono [4]: **pre-processing**, **in-processing** e **post-processing**.

- Le tecniche di **pre-processing**, sulle quali si concentra questa tesi, operano direttamente sui dati di addestramento per rimuovere o ridurre i bias esistenti prima che il modello venga addestrato.
- Le tecniche di **in-processing** modificano l'algoritmo di apprendimento, ad esempio aggiungendo vincoli alla funzione obiettivo per penalizzare soluzioni non eque.
- Le tecniche di **post-processing** agiscono sulle predizioni di un modello già addestrato, aggiustandone le soglie di decisione per bilanciare accuratezza e fairness [9].

2.2.3 Principali Metriche di Fairness di Gruppo

Per quantificare la fairness di gruppo, si utilizzano diverse metriche, spesso in conflitto tra loro. Le più comuni, basate sul concetto di indipendenza statistica o sue varianti, includono:

- **Statistical Parity Difference (SPD)**: Misura la differenza nella probabilità di ottenere un esito favorevole tra il gruppo non privilegiato ($A = 0$) e quello privilegiato ($A = 1$). Un valore ideale è 0 [8].

$$\text{SPD} = P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)$$

- **Disparate Impact (DI)**: È il rapporto tra le stesse probabilità. Un valore ideale è 1. Un valore inferiore a 0.8 è spesso considerato indicativo di discriminazione [8].

$$\text{DI} = \frac{P(\hat{Y} = 1|A = 0)}{P(\hat{Y} = 1|A = 1)}$$

- **Equal Opportunity Difference (EOD)**: Si concentra sull'uguaglianza dei tassi di veri positivi (TPR) tra i gruppi. Il valore ideale è 0 [9].

$$\text{EOD} = \text{TPR}_{A=0} - \text{TPR}_{A=1}$$

- **Average Odds Difference (AOD)**: Estende la EOD considerando la media delle differenze sia del TPR che del False Positive Rate (FPR). Un valore ideale è 0 [9].

$$\text{AOD} = \frac{1}{2}\left[(\text{FPR}_{A=0} - \text{FPR}_{A=1}) + (\text{TPR}_{A=0} - \text{TPR}_{A=1})\right]$$

2.3 Fairness nell'Apprendimento Federato

L'applicazione dei principi di fairness al paradigma del Federated Learning non è una semplice trasposizione, ma introduce un insieme di sfide uniche. Le proprietà intrinseche del FL, come i dati non-IID e i vincoli di privacy, generano due problematiche principali: l'**eterogeneità del bias**, per cui la natura del bias può variare drasticamente da client a client, e l'impossibilità di condividere **statistiche globali** necessarie a molte tecniche di fairness, poiché ciò costituirebbe una fuga di informazioni (*information leakage*).

In risposta a queste sfide, è emerso il campo del *Fair Federated Learning* (FFL) [17]. Tuttavia, la letteratura esistente presenta spesso limitazioni rispetto a uno scenario federato realistico e strettamente privato. Molti approcci, infatti, aggirano il problema della visione locale assumendo la possibilità di condividere statistiche aggregate con il server [12, 16], rilassando di fatto i vincoli di privacy. Altre soluzioni richiedono la presenza di un dataset ausiliario sul server [6], un assunto spesso irrealistico. Una porzione significativa della ricerca, infine, si è concentrata su meccanismi di in-processing e post-processing, lasciando inesplorata una questione fondamentale.

Da questa analisi emerge una lacuna evidente e di grande rilevanza pratica. Mentre le tecniche di pre-processing rappresentano uno standard consolidato in contesti centralizzati, la loro efficacia in un ambiente federato strettamente locale e non-IID rimane un'incognita. La domanda fondamentale a cui la letteratura attuale non fornisce una risposta chiara è la seguente:

> *Come si comportano gli algoritmi di fairness pre-processing classici, quando vengono adattati per operare in totale isolamento su ogni client, senza alcuna forma di comunicazione statistica, e in condizioni di forte eterogeneità del bias?*

Questa tesi si inserisce precisamente in tale lacuna. L'obiettivo non è proporre un nuovo algoritmo di FFL, ma condurre una **rigorosa analisi di benchmark** per comprendere e quantificare le performance, i limiti e i trade-off di tecniche di pre-processing consolidate quando vengono impiegate in uno scenario federato realistico e vincolato. Questo studio mira a fornire una baseline empirica essenziale, oggi mancante, per qualunque ricerca futura in questo dominio.

CAPITOLO 3

Metodo di Ricerca

3.1 Domande di Ricerca

Dall'analisi dello stato dell'arte emerge la necessità di investigare il comportamento di tecniche di fairness pre-processing in ambienti federati. Per indirizzare sistematicamente questa lacuna, lo studio è articolato attorno alle seguenti domande di ricerca.

> **RQ$_1$.** *Qual è l'impatto delle tecniche di preprocessing (*Reweighting, Optimized Preprocessing, Disparate Impact Remover*) sulla **fairness** e sull'**accuratezza** del modello globale federato?*

L'efficacia di queste tecniche è validata in contesti centralizzati, ma la loro applicazione in scenari federati, dove ogni client ha una visione parziale dei dati, non garantisce un risultato analogo sul modello aggregato. Questa domanda mira a stabilire l'impatto fondamentale di tali metodi in un paradigma distribuito, verificando se l'intervento locale si traduca in una modifica significativa delle performance globali del sistema.

RQ$_2$. *Come varia l'efficacia di ciascuna tecnica al variare del grado di eterogeneità dei dati (α)?*

L'eterogeneità dei dati (Non-IID) è una caratteristica intrinseca e critica degli scenari federati. È imperativo analizzare come la performance delle strategie di mitigazione del bias sia modulata da questo fattore. La presente domanda investiga la robustezza di ciascun algoritmo, con l'obiettivo di mapparne l'efficacia in funzione della distribuzione statistica del bias tra i client.

RQ$_3$. *Qual è il trade-off tra fairness e accuratezza per ciascuna tecnica, e quale offre il miglior compromesso in un contesto federato?*

Le metriche di fairness e accuratezza sono spesso in competizione. Questa domanda si propone di quantificare tale trade-off per ogni tecnica analizzata nell'ambiente federato, dove le interazioni tra gli aggiornamenti locali possono generare dinamiche complesse. L'obiettivo è fornire una valutazione comparativa che identifichi l'approccio con il rapporto costo-beneficio più vantaggioso in termini di performance bilanciate.

3.1.1 Reweighting

La tecnica di Reweighting [10] mira a bilanciare il dataset assegnando pesi diversi a ciascuna istanza, in modo da neutralizzare le correlazioni spurie tra l'attributo sensibile e l'etichetta della classe. L'obiettivo è far sì che il dataset pesato soddisfi la parità statistica. Il peso w per un'istanza con attributo sensibile $A = a$ e classe $Y = y$ viene calcolato per rendere le due variabili statisticamente indipendenti:

$$w(a, y) = \frac{P(A = a) \cdot P(Y = y)}{P(A = a, Y = y)}$$

dove le probabilità sono stimate empiricamente sul dataset. Nel nostro contesto federato, **ogni client calcola e applica questi pesi in modo autonomo, basandosi esclusivamente sulle frequenze osservate nel proprio sottoinsieme di dati locale**. Questo rende la tecnica intrinsecamente "federated-friendly" e non richiede alcuna modifica al protocollo FL.

3.1.2 Optimized Preprocessing

L'Optimized Preprocessing [3] è una tecnica che trasforma il dataset attraverso un problema di ottimizzazione. L'obiettivo è trovare una nuova rappresentazione dei dati che minimizzi la distorsione rispetto ai dati originali, pur soddisfacendo vincoli di fairness. . La trasformazione è stata implementata utilizzando la libreria `CVXPY[5]`. Questa applicazione locale testa la capacità dell'algoritmo di funzionare in un'impostazione completamente decentralizzata.

3.1.3 Disparate Impact Remover

Il Disparate Impact Remover (DIR) [8] è una tecnica che modifica i valori delle feature per rimuovere il bias, con l'obiettivo di raggiungere la parità di impatto. L'algoritmo "ripara" i valori delle feature trasformandoli in base al loro rango all'interno dei rispettivi gruppi (privilegiati e non). Per l'implementazione è stata utilizzata la libreria `Fairlearn[14]`, impostando il livello di riparazione al massimo (`repair_level=1.0`). Questa scelta metodologica è cruciale, poiché valuta il comportamento di DIR in condizioni di informazione parziale, un test severo per un algoritmo che si basa su proprietà distributive.

3.2 Dataset Utilizzati

Per la validazione empirica degli algoritmi sono stati selezionati due dataset ampiamente riconosciuti come benchmark nella letteratura sulla fairness. La scelta è ricaduta su di essi poiché rappresentano due distinti domini applicativi e problematiche di bias, consentendo così una valutazione robusta e diversificata delle tecniche analizzate.

3.2.1 COMPAS Dataset

Il primo dataset, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [1], è uno standard de-facto per l'analisi del bias nel sistema di giustizia penale. Contiene dati utilizzati per stimare il rischio di recidiva, con l'obietti-

vo di predire se un individuo commetterà un nuovo reato entro due anni. L'attributo sensibile è la **razza** (race), e i bias razziali ampiamente documentati presenti nei dati lo rendono un banco di prova essenziale per le tecniche di mitigazione.

3.2.2 Adult Dataset

Il secondo dataset, Adult [11], noto anche come Census Income, sposta l'analisi in un contesto socio-economico. L'obiettivo del classificatore è predire se il reddito annuo di un individuo superi la soglia di $50.000, basandosi su una serie di attributi demografici. In questo scenario, l'attributo sensibile è il **genere** (sex), e il dataset è un caso di studio emblematico per l'analisi di disparità economiche storiche, offrendo un contesto di valutazione complementare a quello di COMPAS.

3.3 Simulazione dell'Eterogeneità e del Bias

Per simulare scenari realistici di dati Non-IID, i dataset sono stati partizionati tra i client utilizzando una **distribuzione di Dirichlet** [15] sull'etichetta della classe. Il parametro di concentrazione α controlla il livello di eterogeneità:

- $\alpha = 100$: Bassa eterogeneità (distribuzioni dei dati quasi IID).
- $\alpha = 10$: Eterogeneità moderata.
- $\alpha = 1$: Alta eterogeneità (i dati di ogni client sono fortemente sbilanciati verso poche classi).

Questo partizionamento non solo crea eterogeneità nelle etichette, ma induce anche una **eterogeneità del bias**: la distribuzione dei gruppi protetti e le relative metriche di fairness variano da client a client.

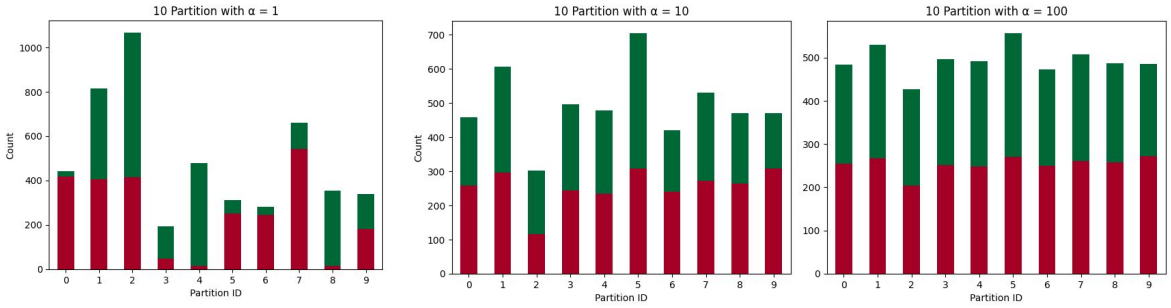

Figura 3.1: Variare delle distribuzioni sul Dataset Compas a diversi livelli di eterogeneità

La Figura 3.1 illustra, attraverso tre pannelli, l'impatto del parametro di concentrazione α della distribuzione di Dirichlet sulla ripartizione dei dati tra 10 client. Ogni pannello rappresenta uno scenario di eterogeneità differente:

- **Pannello di sinistra ($\alpha = 1$):** In questo scenario, si osserva un'elevata eterogeneità. I dati di ciascun client sono fortemente sbilanciati verso uno dei due gruppi protetti, simulando una condizione di *non-IID* (non indipendenti e identicamente distribuiti) estrema.

- **Pannello centrale ($\alpha = 10$):** Con un valore intermedio del parametro, l'eterogeneità è moderata. Le distribuzioni dei dati sui singoli client risultano più bilanciate rispetto al caso precedente, sebbene permangano variazioni significative tra di esse.

- **Pannello di destra ($\alpha = 100$):** In questa configurazione, l'eterogeneità è bassa. Le distribuzioni dei dati tra i client sono molto simili, approssimando uno scenario quasi *IID*.

L'analisi di queste visualizzazioni conferma che valori contenuti di α accentuano la natura *non-IID* e il *bias* di gruppo nella distribuzione dei dati. Al contrario, valori più

elevati di α tendono a mitigare tale effetto, avvicinando la ripartizione dei dati a una distribuzione quasi uniforme e identica per ogni client.

3.4 Setup Sperimentale e Classificatore

La valutazione è stata condotta utilizzando il framework `Flower[2]` con 10 client e 50 round di addestramento. Come classificatore è stato utilizzato un modello di **Regressione Logistica**. La scelta di questo modello non è casuale: essendo un **classificatore lineare**, la sua performance dipende fortemente dall'assunzione che le classi siano linearmente separabili nello spazio delle feature. Questa caratteristica è fondamentale per l'interpretazione dei risultati, in particolare per le tecniche che, come DIR, operano una trasformazione geometrica dello spazio dei dati.

CAPITOLO 4

Risultati e Discussione

In questo capitolo presentiamo e discutiamo i risultati sperimentali ottenuti dall'applicazione delle tecniche di preprocessing in un contesto di Federated Learning. I test sono stati condotti sui dataset `COMPAS` e `Adult`, variando il livello di eterogeneità dei client ($\alpha \in \{1, 10, 100\}$). I risultati completi sono riportati nelle tabelle 4.1 e 4.2, che verranno analizzate nel dettaglio in questa sezione.

Tabella 4.1: Risultati sperimentali sul dataset COMPAS. Le frecce indicano la direzione desiderata per la metrica. I valori in grassetto indicano il risultato migliore per ciascuna metrica all'interno di un blocco di α.

Tecnica	α	**Acc.↑**	**Loss↓**	**SPD↓**	**DI↑**	**EOD↓**	**AOD↓**
Standard	1	0.623	0.656	0.315	0.348	0.306	0.290
Reweighting		0.616	0.670	**0.029**	**0.907**	0.121	0.097
Optimized Preprocessing		**0.628**	**0.652**	0.042	0.851	**0.114**	**0.078**
Disparate Impact Remover		0.576	9.971	0.098	0.522	**0.114**	**0.091**
Standard	10	**0.700**	**0.599**	0.336	0.496	0.305	0.260
Reweighting		0.687	0.613	**0.049**	**0.904**	**0.037**	**0.076**
Optimized Preprocessing		0.693	0.610	0.060	0.888	0.108	0.104
Disparate Impact Remover		0.615	8.263	0.144	0.489	0.244	0.159
Standard	100	**0.646**	0.650	0.273	0.544	0.258	0.252
Reweighting		0.631	0.652	0.085	**0.825**	0.133	0.091
Optimized Preprocessing		0.619	**0.635**	0.107	0.783	0.170	0.148
Disparate Impact Remover		0.588	11.14	**0.080**	0.626	**0.105**	**0.076**

Tabella 4.2: Risultati sperimentali sul dataset Adult. Le frecce indicano la direzione desiderata per la metrica. I valori in grassetto indicano il risultato migliore per ciascuna metrica all'interno di un blocco di α.

Tecnica	α	**Acc.↑**	**Loss↓**	**SPD↓**	**DI↑**	**EOD↓**	**AOD↓**
Standard	1	0.700	0.611	0.279	0.385	0.031	0.151
Reweighting		0.709	0.624	0.169	0.572	0.008	0.087
Optimized Preprocessing		0.722	0.517	**0.058**	**0.823**	**0.000**	**0.029**
Disparate Impact Remover		**0.862**	**0.363**	0.065	0.601	0.009	0.031
Standard	10	**0.830**	**0.346**	0.174	0.347	**0.106**	0.094
Reweighting		0.824	0.364	0.071	0.686	0.146	**0.080**
Optimized Preprocessing		0.813	0.414	**0.028**	**0.870**	0.244	0.145
Disparate Impact Remover		0.798	0.680	0.108	0.513	0.132	0.084
Standard	100	**0.852**	**0.321**	0.163	0.390	0.059	0.069
Reweighting		0.845	0.340	0.069	0.686	0.196	0.111
Optimized Preprocessing		0.834	0.378	**0.042**	**0.799**	0.272	0.148
Disparate Impact Remover		0.824	0.569	0.116	0.474	**0.052**	**0.051**

4.1 Discussione dei Risultati

RQ$_1$. *Qual è l'impatto delle tecniche di preprocessing (*Reweighting, Optimized Preprocessing, Disparate Impact Remover*) sulla* ***fairness*** *e sull'accuracy nei modelli federati?*

L'introduzione di tecniche di preprocessing in un ambiente di apprendimento federato ha un impatto profondo e sfaccettato sia sulla fairness che sull'accuracy del modello aggregato. I risultati sperimentali mostrano che, in generale, queste tecniche riescono nel loro intento primario di mitigare il bias, ma con effetti collaterali e performance variabili a seconda della tecnica e del contesto.

Impatto sulla Fairness.

- **Miglioramento significativo del bias:** Tutte e tre le tecniche di preprocessing testate hanno dimostrato di poter migliorare in modo sostanziale le metriche di fairness rispetto al modello *Standard* (baseline). Ad esempio, come riportato nella Tabella 4.1, sul dataset COMPAS con $\alpha = 1$, *Reweighting* riduce la *Statistical Parity Difference* (SPD) da 0.315 a **0.029**. Analogamente, sul dataset Adult ($\alpha = 1$), la Tabella 4.2 mostra che *Optimized Preprocessing* ottiene risultati di fairness quasi perfetti, con una *Equal Opportunity Difference* (EOD) pari a **0.000**.
- **Specializzazione delle tecniche:** Le tecniche non sono universalmente efficaci su tutte le metriche di fairness. *Reweighting* e *Optimized Preprocessing* si dimostrano particolarmente potenti nel migliorare la fairness di gruppo (SPD e *Disparate Impact* - DI). *Disparate Impact Remover*, d'altro canto, pur essendo meno consistente, in alcuni scenari (es. Adult con $\alpha = 100$, Tabella 4.2) eccelle nel ridurre le disparità legate alle opportunità (EOD e *Average Odds Difference* - AOD).

Impatto sull'Accuracy e sulla Stabilità del Modello.

- **Generalmente un calo di accuracy:** Nella maggior parte degli scenari, l'applicazione di una tecnica di fairness introduce un calo, solitamente lieve, dell'accuratezza. Questo rappresenta il classico trade-off. Ad esempio, su Adult

con $\alpha = 100$ (Tabella 4.2), il modello *Standard* raggiunge un'accuracy di 0.852, mentre le altre tecniche si assestano su valori leggermente inferiori.

- **La rottura del trade-off:** Il risultato più notevole è quello di *Disparate Impact Remover* sul dataset Adult con alta eterogeneità ($\alpha = 1$), visibile nella Tabella 4.2. In questo caso specifico, la tecnica non solo migliora la fairness, ma **aumenta drasticamente l'accuratezza**, passando dallo 0.700 del modello standard a **0.862**. Questo dimostra che il trade-off accuratezza-fairness non è una legge universale e che, in determinate condizioni, la mitigazione del bias può anche migliorare le performance predittive.

- **Impatto sulla convergenza (Loss):** Un aspetto critico è l'impatto sulla stabilità del training. Come si osserva nella Tabella 4.1, *Disparate Impact Remover* sul dataset COMPAS ha prodotto valori di **loss estremamente elevati** (es. 11.14 vs 0.650 del modello *Standard* con $\alpha = 100$), indicando gravi problemi di convergenza. Ciò suggerisce che la trasformazione operata da questa tecnica può rendere i dati molto difficili da apprendere per il modello, rendendola di fatto inutilizzabile in quel contesto.

Risposta alla Domanda di Ricerca 2

RQ$_2$. *Come varia l'efficacia di ciascuna tecnica al cambiare del grado di eterogeneità dei dati α?*

Il grado di eterogeneità dei dati tra i client, controllato dal parametro α, è un fattore determinante per l'efficacia delle tecniche di preprocessing. L'analisi delle tabelle 4.1 e 4.2 mostra che non solo le performance generali, ma anche la classifica relativa delle tecniche cambia drasticamente al variare di α.

- **Standard (Baseline):** Sul dataset Adult, il modello *Standard* segue il comportamento teorico atteso: l'accuratezza aumenta al diminuire dell'eterogeneità (all'aumentare di α). Sul dataset COMPAS, invece, si osserva un comportamento anomalo, con le performance migliori per un valore intermedio ($\alpha = 10$, Tabella 4.1), suggerendo una complessa interazione tra la distribuzione dei dati e la natura del problema.

- **Reweighting:** Questa tecnica si dimostra **la più robusta e stabile** al variare di α. Fornisce un miglioramento della fairness consistente e prevedibile su entrambi i dataset, indipendentemente dal livello di eterogeneità. La sua efficacia non sembra dipendere in modo critico da α, rendendola una scelta affidabile in scenari in cui l'eterogeneità è sconosciuta o variabile.

- **Optimized Preprocessing:** La sua efficacia è più sensibile a α. Sul dataset Adult (Tabella 4.2), eccelle in condizioni di alta eterogeneità ($\alpha = 1$), ottenendo la migliore fairness complessiva. Tuttavia, al diminuire dell'eterogeneità ($\alpha \geq 10$), la sua capacità di mitigare il bias su EOD e AOD si riduce notevolmente, pur mantenendo ottimi risultati su SPD e DI.

- **Disparate Impact Remover:** È la tecnica **più sensibile in assoluto** all'eterogeneità dei dati, con un comportamento quasi dicotomico:

  - **Su Adult:** Come mostrato nella Tabella 4.2, è la tecnica migliore in assoluto in condizioni di **alta eterogeneità** ($\alpha = 1$), superando tutte le altre in accuratezza e mantenendo un'ottima fairness. La sua efficacia sull'accuracy, però, cala all'aumentare di α.
  - **Su COMPAS:** È inefficace e problematica (loss alta, accuracy bassa) **per tutti i valori di α testati**, come documentato nella Tabella 4.1.

  Questa forte dipendenza dal contesto (dataset + eterogeneità) la rende una tecnica ad alto rischio, che può portare a risultati eccezionali o fallimentari.

Risposta alla Domanda di Ricerca 3

RQ$_3$. *Qual è il trade-off tra fairness e accuracy per ciascuna tecnica di preprocessing, e quale bilanciamento ottiene il miglior compromesso?*

L'analisi del trade-off tra fairness e accuracy, basata sui dati delle tabelle 4.1 e 4.2, rivela che non esiste una risposta unica, ma il "miglior compromesso" dipende dagli obiettivi specifici, dal contesto operativo e dalla tolleranza al rischio.

- **Reweighting:** Offre il **compromesso più equilibrato e affidabile**. Su entrambi i dataset e a tutti i livelli di α, "paga" un prezzo molto piccolo in termini di

accuratezza per "acquistare" un miglioramento molto grande della fairness (specialmente su SPD e DI). Se l'obiettivo è un miglioramento della fairness significativo con un impatto minimo e prevedibile sulle performance, *Reweighting* rappresenta la scelta ottimale.

- **Optimized Preprocessing:** Fornisce un ottimo bilanciamento, spesso raggiungendo i **picchi più alti di fairness di gruppo (SPD/DI)**. Il suo trade-off è vantaggioso, ma può nascondere un compromesso secondario: nella Tabella 4.2 si nota che, al diminuire dell'eterogeneità, ottimizza SPD/DI a scapito di EOD/AOD. È la scelta ideale se la parità statistica è l'obiettivo primario e si accetta una potenziale debolezza su altre metriche di fairness.

- **Disparate Impact Remover:** Rappresenta l'approccio più estremo e volatile.

  - Nella maggior parte dei casi (tutti gli scenari su COMPAS, Tabella 4.1), il suo trade-off è **estremamente svantaggioso**: sacrifica enormemente l'accuratezza e la stabilità del modello per ottenere miglioramenti di fairness, rendendolo impraticabile.

  - Nello scenario specifico di Adult con $\alpha = 1$ (Tabella 4.2), **rompe il trade-off**, ottenendo il miglior risultato sia in accuracy che in fairness. In questo caso, è indiscutibilmente la scelta migliore.

  - In altri scenari (Adult, $\alpha \geq 10$), offre un compromesso diverso: sacrifica più accuracy rispetto alle altre tecniche per ottenere i migliori risultati su EOD e AOD (cfr. Tabella 4.2).

CAPITOLO 5

Conclusioni e Sviluppi Futuri

Questo capitolo finale riassume i contributi della tesi, ne riconosce i limiti e propone diverse direzioni promettenti per la ricerca futura, consolidando il percorso svolto e guardando alle sfide che attendono.

5.1 Sintesi dei Risultati Principali

L'analisi sperimentale ha fornito risposte complesse alle domande di ricerca iniziali, evidenziando dinamiche non sempre prevedibili. I risultati chiave sono:

- **Trade-off Accertato ma non Assoluto:** L'applicazione di tecniche di pre-processing migliora l'equità, ma spesso a un costo in termini di accuratezza. Tuttavia, questo trade-off non è una legge inviolabile. La tecnica *Disparate Impact Remover*, in condizioni di alta eterogeneità, ha migliorato simultaneamente accuratezza e fairness, dimostrando che la mitigazione del bias può, in casi specifici, agire in sinergia con l'obiettivo predittivo.

- **Impatto Critico dell'Eterogeneità:** L'eterogeneità dei dati è emersa come un fattore determinante. Non solo influenza le performance globali, ma modifica

la classifica di efficacia degli algoritmi, rendendo la scelta della tecnica ottimale fortemente dipendente dal contesto.

- **Stabilità del Reweighting vs. Volatilità del DIR:** Tra le tecniche analizzate, *Reweighting* si è distinta come la soluzione più stabile e bilanciata, offrendo un miglioramento della fairness prevedibile a fronte di un costo in accuratezza contenuto. Al contrario, *Disparate Impact Remover* ha mostrato un comportamento estremamente volatile, con performance eccezionali in uno scenario e un fallimento catastrofico in un altro.

5.2 Implicazioni dello Studio e Contesto Scientifico

I risultati di questa tesi hanno implicazioni che vanno oltre la mera analisi sperimentale, inserendosi in un dibattito scientifico molto attivo.

- **Per la Pratica:** La lezione più importante è che non esiste una "soluzione universale" per la fairness in ambito federato. Gli sviluppatori devono condurre validazioni mirate, tenendo conto del proprio caso d'uso, del livello di eterogeneità e della specifica nozione di fairness da privilegiare.
- **Per la Ricerca:** Questo studio fornisce una forte **motivazione empirica per superare i limiti degli approcci pre-processing statici**. I comportamenti anomali del DIR, in particolare, evidenziano un'interazione complessa tra trasformazione dei dati, distribuzione e apprendimento distribuito. Le nostre scoperte suggeriscono che la rigidità del pre-processing, che applica una trasformazione fissa prima del training, è intrinsecamente fragile in ambienti eterogenei. Questo rafforza la tesi della comunità scientifica che si sta orientando verso approcci *in-processing* (che intervengono durante il training) e *post-processing* (che correggono le predizioni), in quanto potenzialmente più flessibili e adattivi.

5.3 Limitazioni e Validità dei Risultati

È fondamentale riconoscere i confini di questo studio per una corretta interpretazione dei risultati.

- **Generalizzabilità:** L'analisi è stata condotta su due soli dataset di benchmark. Scenari reali potrebbero presentare caratteristiche differenti che potrebbero influenzare i risultati.
- **Portata degli Algoritmi:** Lo studio si è focalizzato esclusivamente su tecniche di pre-processing. Un confronto con metodi *in-processing* e *post-processing* è un passo necessario per una visione completa.
- **Specificità della Configurazione:** I risultati sono legati alla configurazione sperimentale adottata (es. regressione logistica, FedAvg). Architetture più complesse o algoritmi di aggregazione diversi potrebbero modificare gli equilibri osservati.

5.4 Lavori Futuri e Prospettive di Ricerca

Le scoperte di questa tesi aprono la strada a diverse direzioni di ricerca specifiche e promettenti, nate direttamente dalle questioni rimaste aperte.

- **Verifica Empirica con Modelli Non-Lineari:** La nostra ipotesi principale sulla volatilità di *Disparate Impact Remover* deve essere testata. Un lavoro futuro prioritario consiste nel **ripetere gli esperimenti sostituendo la regressione logistica con classificatori non lineari** (es. Support Vector Machine con kernel RBF o reti neurali) per investigare se un modello più espressivo sia più robusto alle distorsioni geometriche dei dati.
- **Confronto con Paradigmi Alternativi:** I limiti del pre-processing emersi in questo studio motivano un confronto diretto con approcci più recenti. Si propone uno **studio comparativo con un framework di post-processing federato**, dove ogni client calibra localmente il modello globale. Tale studio quantificherebbe i vantaggi di un approccio più flessibile e decentralizzato.
- **Ottimizzazione della Fairness tramite Condivisione di Metriche Aggregate**: Una promettente evoluzione di questo lavoro consisterebbe nell'introdurre un paradigma in cui i client condividono metriche di fairness con un server centrale. Invece di operare in isolamento, i client trasmetterebbero indicatori anonimi

e aggregati relativi all'equità del sistema. Questo consentirebbe al server di ottenere una visione olistica e globale, abilitando strategie di mitigazione del bias (ad esempio, attraverso la ricalibrazione del modello) significativamente più informate e precise rispetto all'approccio puramente locale analizzato in questa tesi.

- **Sviluppo di Approcci Ibridi e Adattivi:** Una frontiera di ricerca promettente è la progettazione di metodi ibridi che combinino diverse strategie (es. pre-processing locale e in-processing a livello server) o di sistemi adattivi capaci di selezionare dinamicamente la tecnica di fairness più adatta in base a una stima in tempo reale dell'eterogeneità del sistema.

In conclusione, questo lavoro di tesi ha contribuito a illuminare la complessa relazione tra equità, accuratezza ed eterogeneità dei dati nell'apprendimento federato. I risultati non solo forniscono una guida pratica, ma, soprattutto, aprono nuove e stimolanti questioni di ricerca. L'auspicio è che questo studio possa servire come base per lo sviluppo futuro di sistemi di intelligenza artificiale distribuiti che siano non solo performanti, ma anche e soprattutto equi e affidabili.

Bibliografia

[1] Julia Angwin et al. “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks”. In: *ProPublica* 23.2016 (2016), pp. 139–159.

[2] Daniel J. Beutel et al. “Flower: A Friendly Federated Learning Research Framework”. In: *Proceedings of the 1st Workshop on Distributed Machine Learning*. 2020. URL: https://arxiv.org/abs/2007.14390.

[3] Flavio Calmon et al. “Optimized pre-processing for discrimination prevention”. In: *Advances in neural information processing systems* 30 (2017).

[4] Simon Caton e Christian Haas. “Fairness in Machine Learning: A Survey”. In: *arXiv preprint arXiv:2010.04053* (2020).

[5] Steven Diamond e Stephen Boyd. “CVXPY: A Python-embedded modeling language for convex optimization”. In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[6] Lingjuan Du et al. “Fairness-aware agnostic federated learning”. In: *2021 IEEE International conference on big data (Big Data)*. IEEE. 2021, pp. 101–110.

[7] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–221.

[8] Michael Feldman et al. “Certifying and removing disparate impact”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 259–268.

[9] Moritz Hardt, Eric Price e Nati Srebro. “Equality of opportunity in supervised learning”. In: *Advances in neural information processing systems*. Vol. 29. 2016.

[10] Faisal Kamiran e Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33.

[11] Ron Kohavi. “Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid”. In: *Proceedings of the second international conference on knowledge discovery and data mining* (1996), pp. 202–207.

[12] Teng Li et al. “FairFed: A General Fair Federated Learning Framework”. In: *arXiv preprint arXiv:1905.12224v3*. 2020.

[13] Brendan McMahan et al. “Communication-Efficient Learning of Deep Networks from Decentralized Data”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. A cura di Aarti Singh e Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[14] Hilde Weerts et al. “Fairlearn: Assessing and Improving Fairness of AI Systems”. In: *Journal of Machine Learning Research* 24.188 (2023), pp. 1–8.

[15] Mikhail Yurochkin et al. *Bayesian Nonparametric Federated Learning of Neural Networks*. 2019. arXiv: 1905.12022 [stat.ML]. URL: https://arxiv.org/abs/1905.12022.

[16] Jinyuan Zeng et al. “Improving fairness and privacy in federated learning”. In: *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2021, pp. 889–898.

[17] Lingjuan Zhang e Lixu Wang. “Fairness in federated learning: A survey”. In: *arXiv preprint arXiv:2208.13401* (2022).

[18] Yue Zhao et al. "Federated learning with non-iid data". In: *arXiv preprint arXiv:1806.00582* (2018).