UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Mitigazione del Bias nei Dati Attraverso Refactoring Supportato da Large Language Models

RELATORE

Prof. Fabio Palomba

Dott. Gianmario Voria

Università degli Studi di Salerno

CANDIDATO

Luigi Guida

Matricola: 0512114254

Anno Accademico 2023-2024

Questa tesi è stata realizzata nel

Injustice anywhere is a threat to justice everywhere.

Martin Luther King Jr.

Abstract

Questa tesi esplora il tema della fairness nel machine learning, con particolare attenzione alla rilevazione e correzione di bias nei dataset tramite l'uso di Large Language Models (LLM). L'equità nei modelli di machine learning è essenziale per evitare decisioni discriminatorie che potrebbero influenzare negativamente individui o gruppi sociali. Lo studio si concentra sull'identificazione dei sintomi di fairness e analizza come i LLM possano essere utilizzati non solo per calcolare e interpretare tali sintomi, ma anche per eseguire refactoring sui dati al fine di ridurre il bias. Attraverso esperimenti condotti su diversi dataset, la tesi dimostra che l'applicazione di tecniche di resampling e altre strategie di pre-processing suggerite dai LLM possono migliorare i livelli di fairness, riducendo le disparità tra gruppi privilegiati e non privilegiati. I risultati ottenuti evidenziano le potenzialità e i limiti dell'approccio proposto, sottolineando l'importanza di ulteriori ricerche per rendere i modelli di machine learning sempre più equi e imparziali.

Indice

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Contesto Applicativo

La crescente adozione dei modelli di machine learning in settori critici, come la finanza, la sanità e la giustizia, ha portato alla luce problematiche legate alla fairness e all'impatto sociale delle decisioni automatizzate. I sistemi di machine learning possono presentare bias intrinseci, derivanti dai dati con cui vengono addestrati o da errori di progettazione, che portano a decisioni discriminatorie nei confronti di gruppi sociali o demografici. Questo bias può risultare in disparità ingiustificate nelle predizioni e nell'equità del modello, compromettendo la fiducia nei sistemi automatizzati e amplificando le disuguaglianze esistenti. Nel tentativo di affrontare queste problematiche, l'uso di Large Language Models sta emergendo come uno strumento promettente per rilevare e mitigare il bias nei dati. Gli LLM, grazie alla loro capacità di elaborare grandi volumi di dati e comprendere complessi contesti linguistici, sono in grado di identificare sintomi di unfairness e proporre soluzioni di refactoring dei dati per ridurre il bias prima che i dati vengano utilizzati per l'addestramento dei modelli.

1.2 Motivazioni e Obiettivi

Il bias nei modelli di machine learning non è solo un problema tecnico, ma ha implicazioni etiche e sociali profonde, poiché le decisioni automatizzate possono influenzare direttamente la vita delle persone. Garantire che i modelli siano equi e non discriminatori è cruciale per evitare conseguenze negative e per costruire sistemi di decision-making che riflettano l'integrità e la giustizia. La motivazione principale di questo lavoro nasce dall'esigenza di sviluppare tecniche efficaci per mitigare il bias nei dati, migliorando così l'equità nei modelli di machine learning. L'obiettivo di questa tesi è esplorare il ruolo dei LLM nella mitigazione del bias, focalizzandosi sulla fase di pre-processing dei dati. Si indaga in che misura gli LLM siano in grado di identificare i sintomi di unfairness nei dataset e di proporre strategie di refactoring che riducano il bias, garantendo un trattamento equo dei dati prima dell'addestramento dei modelli. Questo studio si propone di valutare l'efficacia degli LLM in questo contesto, aprendo la strada a sviluppi futuri in cui il loro impiego possa estendersi anche alle fasi successive del processo di machine learning.

1.3 Risultati Ottenuti

I risultati di questo studio hanno confermato che i Large Language Models (LLM) sono in grado di identificare i principali sintomi di bias nei dataset, contribuendo così a migliorare l'equità nei modelli di machine learning. Attraverso l'uso di tecniche di refactoring dei dati, gli LLM hanno dimostrato di poter ridurre il bias prima dell'addestramento dei modelli, specialmente in contesti con strutture di dati meno complesse. Nonostante i risultati positivi, l'efficacia degli LLM nel refactoring dipende dall'accuratezza del contesto fornito, e in alcuni casi è stato necessario un intervento umano per guidare il modello verso soluzioni ottimali. In generale, gli LLM si sono dimostrati strumenti promettenti per migliorare la fairness, ma ulteriori ricerche saranno necessarie per ampliare il loro utilizzo a fasi più avanzate del processo di machine learning.

1.4 Struttura della Tesi

In questa tesi verrà esaminato come i LLM possano essere utilizzati per mitigare il *bias* nei dati, con un focus particolare sull'identificazione dei sintomi di *unfairness*. Il documento è strutturato come segue:

- **Capitolo 2: Background e Stato dell'Arte**
  In questo capitolo verranno introdotti i concetti fondamentali di fairness nel machine learning, con una panoramica sui principali tipi di bias che possono influenzare i dati e i modelli. Saranno inoltre discusse le metriche utilizzate per misurare la fairness e le principali strategie per mitigare il bias, incluse le fasi di pre-processing, in-processing e post-processing.

- **Capitolo 3: Metodologia Sperimentale**
  Qui verranno descritti nel dettaglio gli strumenti utilizzati, come GPT-4, e le tecniche di prompting adottate per ottenere risultati accurati. Saranno presentati i metodi di refactoring suggeriti dai LLM e le strategie di intervento sui dataset, come il resampling e l'oversampling sintetico.

- **Capitolo 4: Analisi dei Risultati**
  In questo capitolo verranno analizzati i risultati degli esperimenti, confrontando i sintomi di fairness prima e dopo l'applicazione delle tecniche di refactoring. Verrà esaminata l'efficacia dei LLM nell'identificare e mitigare il bias nei dataset scelti.

- **Capitolo 5: Conclusioni**
  Infine, verranno tratte le conclusioni sul lavoro svolto, con una discussione sui limiti dell'approccio proposto e sulle direzioni future per migliorare ulteriormente l'equità nei sistemi di machine learning.

CAPITOLO 2

Background e Stato dell'Arte

Questo capitolo fornisce una panoramica, basata sullo stato dell'arte e sui lavori esistenti, dei concetti chiave affrontati in questo studio. In particolare, verrà analizzata la fairness nel contesto del machine learning, approfondendo le principali metriche utilizzate per valutare l'equità e l'affidabilità dei modelli. Sarà inoltre delineata una distinzione tra i diversi tipi di bias che possono influenzare sia i dataset che il comportamento dei modelli, esaminando le fasi dello sviluppo in cui è possibile identificare e mitigare tali bias. Infine, verranno illustrati i sintomi della fairness, evidenziando i segnali di ingiustizia nei dati, insieme a una descrizione dettagliata dei dataset utilizzati in questo studio.

2.1 Fairness

L'argomento della fairness nel machine learning sta assumendo un ruolo sempre più centrale nello sviluppo dei modelli di decisione automatizzata ai giorni d'oggi. Non è da sottovalutare l'impatto negativo a livello sociale che può essere causato da prodotti iniqui e discriminatori.[1] Ci sono stati vari casi particolarmente influenti che hanno, nel corso degli anni, spinto i ricercatori ad approfondire gli studi inerenti alla fairness nel contesto del machine learning. Per esempio, uno strumento di assunzione di Amazon era prevenuto contro le donne, gli strumenti di elaborazione del linguaggio sono più accurati in inglese scritto da anglosassoni che da persone di altre razze. Quindi, come le persone, gli algoritmi sono vulneribili ai pregiudizi, rendendo le loro decisioni "unfair". Nel contesto del machine learning, la fairness è l'assenza di qualsiasi pregiudizio o favoritismo nei confronti di un individuo o di un gruppo in base alle sue caratteristiche intrinseche o acquisite. Pertanto, un modello unfair è quello le cui decisioni sono distorte verso un particolare gruppo di persone.[2]

2.1.1 Gruppi privilegiati e non privilegiati

Un aspetto cruciale nella valutazione della fairness nei modelli di machine learning riguarda la distinzione tra gruppi privilegiati e non privilegiati. Questi concetti sono fondamentali poiché riflettono le disparità di trattamento che possono emergere a causa di bias presenti nei dati o nei modelli stessi. Per gruppi privilegiati e non privilegiati si intendono gruppi che, per via di bias all'interno dei dati e dei modelli di machine learning, ricevono dei trattamenti favorevoli o sfavorevoli.[3] Per trattamento favorevole intendiamo che il gruppo privilegiato è, in modo sproporzionato, più probabile di essere classificato positivamente. Viceversa, per i gruppi non privilegiati, è molto più elevata la probabilità di essere classificato negativamente.[4]

2.1.2 Attributi protetti

La suddivisione degli elementi all'interno del dataset in gruppi con disparità di privilegi è dettata principalmente dalla presenza di attributi protetti nel set di dati.

Per attributi protetti si intende caratteristiche personali presenti in un dataset che rappresentano dei gruppi sociali o demografici specifici, i quali sono protetti dalla discriminazione in base a leggi o principi etici. Sono delle caratteristiche che non dovrebbero essere utilizzate come basi per le decisioni di un modello di machine learning. Alcuni esempi sono genere, etnia ed età.[4]

2.1.3 Metriche di fairness

Quando si inizia a lavorare per garantire la fairness nel machine learning è bene avere a disposizione degli strumenti per misurare e valutare quanto un modello di machine learning tratti in modo equo i vari gruppi all'interno di un set di dati. Per questo vengono introdotte le metriche di fairness, atte a individuare e mitigare bias che potrebbero portare a risultati unfair. Esistono dozzine di metriche che, a conti fatti, altro non fanno che andare a "riscaldare una vecchia minestra".[1] Quindi è possibile andare a limitare il numero di metriche da andare ad utilizzare, in modo da non rendere il tutto inutilmente complesso e confusionario. In questo lavoro, tra le varie metriche, ne verranno utilizzate 3 in particolar modo: Statistical Parity, Equalized Odds e Disparate Impact.

Statistical/Demographic Parity

Questa metrica definisce la fairness come un'uguale probabilità nell'essere classificati positivamente tra i gruppi privilegiati e non privilegiati. Presenta tuttavia uno svantaggio, in quanto non vengono prese in considerazione le differenze tra i gruppi.[4]

$$Pr(\hat{y} = 1 \mid g_i) = Pr(\hat{y} = 1 \mid g_j)$$

Equalized Odds

Questa metrica richiede che il modello abbia tassi di veri positivi (TPR) e falsi positivi (FPR) simili per tutti i gruppi, in modo da non favorire dei gruppi rispetto ad altri in termini di precisione delle predizioni.

$$Pr(\hat{y} = 1 \mid y = 1 \& g_i) = Pr(\hat{y} = 1 \mid y = 1 \& g_j) \ \&$$

$$Pr(\hat{y} = 1 \mid y = 0 \& g_i) = Pr(\hat{y} = 1 \mid y = 0 \& g_j)$$

Disparate Impact

Simile alla Statistical/Demographic Parity, ma differisce da essa in quanto considera il rapporto tra i gruppi privilegiati e non privilegiati. Va a verificare quindi se la probabilità che un gruppo riceva un outcome positivo risulta sproporzionato tra gruppi privilegiati e non privilegiati.[4]

$$\frac{Pr(\hat{y} = 1 | g_i)}{Pr(\hat{y} = 1 | g_j)}$$

2.1.4 Strategie di fairness nel machine learning

Sono varie le fasi nelle quali è possibile attuare delle strategie per garantire la fairness e mitigare il bias durante lo sviluppo di un modello di machine learning, e vengono suddivise in: pre-processing, in-processing e post-processing. Successivamente verranno descritte nel dettaglio in cosa differiscono queste strategie.[4]

Pre-processing

Gli approcci di pre-processing riconoscono che spesso il problema risiede nei dati stessi, dove le distribuzioni di variabili sensibili o protette possono essere distorte, discriminatorie e/o sbilanciate. Se l'algoritmo è autorizzato a modificare i dati di addestramento, allora è possibile utilizzare un approccio pre-processing.[2] Pertanto, gli approcci di pre-processing tendono a modificare le distribuzioni campionarie delle variabili protette o, più in generale, a eseguire trasformazioni specifiche sui dati con l'obiettivo di eliminare la discriminazione dai dati di addestramento. L'idea principale è quella di addestrare un modello su un dataset "corretto". Il

pre-processing è considerato la parte più flessibile della pipeline di data science, poiché non fa assunzioni riguardo alla tecnica di modellazione che verrà applicata successivamente.[4]

In-processing

Gli approcci di in-processing riconoscono che le tecniche di modellazione possono spesso diventare parziali a causa di caratteristiche dominanti, altri effetti di distribuzione, o perché cercano di bilanciare più obiettivi del modello, ad esempio ottenere un modello che sia sia accurato che equo. Gli approcci di in-processing affrontano questo problema spesso incorporando una o più metriche di equità nelle funzioni di ottimizzazione del modello, nel tentativo di convergere verso una parametrizzazione che massimizzi sia le prestazioni che l'equità. Se è permesso di poter cambiare la modalità di apprendimento, allora l'approccio in-processing può essere utilizzato durante l'addestramento di un modello.[2, 4]

Post-processing

Gli approcci di post-processing riconoscono che il risultato effettivo di un modello di machine learning può essere ingiusto nei confronti di una o più variabili protette e/o sottogruppi all'interno di queste variabili. Pertanto, gli approcci di post-processing tendono ad applicare trasformazioni ai risultati del modello per migliorare l'equità delle previsioni. Il post-processing è uno degli approcci più flessibili, poiché richiede solo l'accesso alle previsioni e alle informazioni sugli attributi sensibili, senza necessità di accedere agli algoritmi o ai modelli di machine learning effettivi. Se l'algoritmo può trattare il modello solo come una black-box senza alcuna capacità di modificare i dati di addestramento o di apprendimento, allora può essere utilizzata solo la post-processing in cui le etichette assegnate dal modello black-box vengono inizialmente riassegnate in base a una funzione durante la fase di post-processing.[2, 4]

2.2 Bias nei dati

I dati sono strettamente associati alla funzionalità di algoritmi e sistemi di intelligenza artificiale. Nei casi in cui i dati di addestramento contengono distorsioni, gli algoritmi addestrati su di essi apprenderanno queste distorsioni e le rifletteranno nelle loro previsioni. Di conseguenza, le distorsioni esistenti nei dati possono influenzare gli algoritmi che utilizzano li utilizzano, producendo risultati distorti. Gli algoritmi possono persino amplificare e perpetuare i pregiudizi esistenti nei dati. Esistono due tipi di bias che sono strettamente correlati ai dati: Data to Algorithm e User to Data.[2]

2.2.1 Data to algorithm

Questo tipo di bias si riferisce alle distorsioni che vengono trasferite dai dati utilizzati per addestrare un modello agli algoritmi stessi. Se i dati di addestramento sono sbilanciati o contengono pregiudizi impliciti o espliciti, questi possono essere incorporati nel modello di machine learning, influenzandone le previsioni e le decisioni.

Measurement Bias

Il measurement bias è un tipo di distorsione che si verifica nel momento in cui gli strumenti o le metodologie utilizzate per raccogliere i dati introducono errori sistematici o pregiudizi nelle misurazioni stesse. Le cause che possono portare ad un errore di misurazione nella raccolta dei dati sono molteplici, tra cui l'utilizzo di strumenti di misurazione inadeguati o difettosi, errori da parte dell'utente, nelle condizioni di misurazione non adatte e così via.

Omitted Variable Bias

Questo tipo di bias occorre nel momento in cui una o più variabili rilevanti vengono trascurate durante lo sviluppo del modello.[2] Può portare quindi a stime distorte da parte del modello, in quanto esso non terrà in considerazione l'influenza che avrebbe potuto avere la variabile omessa nella decisione finale. Occorre principalmente

con variabili omesse che sono correlate a una o più variabili incluse nel modello, o quando influenzano direttamente la variabile dipendente.

Representation Bias

Il bias di rappresentazione deriva dal modo in cui campionamo da una popolazione durante il processo di raccolta dei dati.[2] Può emergere quando alcune caratteristiche, gruppi, o categorie sono sotto-rappresentati o sovra-rappresentati nel dataset, portando a modelli inaccurati, ingiusti o non generalizzabili.

Aggregation Bias

Il bias di aggregazione sorge quando si traggono false conclusioni sugli individui dall'osservazione dell'intera popolazione.[2] In particolar modo, si può incombere in questo tipo di bias quando non si vanno a considerare le differenze tra i vari sottogruppi, andando ad aggregare insieme i dati provenienti da quest'ultimi. L'aggregation bias può essere un esempio del paradosso di Simpson, in cui una tendenza osservata in sottogruppi separati si inverte quando i dati vengono aggregati. Questo avviene perché l'aggregazione maschera le relazioni tra le variabili all'interno dei sottogruppi

Sampling Bias

Questo tipo di bias è simile al representation bias, con la differenza che esso sorge nel momento in cui non viene effettuato un campionamento casuale dei gruppi, facendo quindi in modo che il campione di dati raccolto per un'analisi non rappresenta adeguatamente la popolazione dalla quale è stato estratto. Questo può portare a risultati fuorvianti, poiché le conclusioni tratte dal campione non possono essere generalizzate all'intera popolazione.[2]

Longitudinal Data Fallacy

Questo tipo di bias è un errore interpretativo che si verifica quando si analizzano dati raccolti su un lungo periodo di tempo (dati longitudinali) e si traggono conclusioni fuorvianti a causa di una mancata considerazione di alcuni fattori cruciali.

Questa fallacia può derivare da un'errata interpretazione dei cambiamenti nel tempo, dall'ignorare variazioni individuali o da errori nell'aggregazione dei dati.

Linking Bias

Il linking bias si verifica quando i network attributes ottenuti dalle connessioni, dalle attività o dalle interazioni degli utenti differiscono e rappresentano in modo errato il vero comportamento degli utenti.[2] Questo bias può influenzare la costruzione, l'interpretazione e l'analisi delle reti, portando a conclusioni fuorvianti riguardo alle relazioni o alla struttura della rete stessa.

2.2.2 User to Data

Molto spesso i dati utilizzati per l'addestramento di modelli di machine learning sono user-generated. Di conseguenza, eventuali pregiudizi intrinseci negli utenti che vengono, a loro volta, riversati sui dati da loro forniti, possono andare ad influenzare le decisioni di un modello.

Historical Bias

L'historical bias è il pregiudizio e i problemi socio-tecnici già esistenti e può sorgere dal processo di generazione di dati anche con un ottimo campionamento e selezione delle caratteristiche.[2] Ad esempio, se i dati storici mostrano una rappresentanza sproporzionata di uomini rispetto alle donne in un determinato ruolo, i modelli di machine learning addestrati su questi dati possono perpetuare o addirittura amplificare questi squilibri.

Population Bias

Questo tipo di bias si verifica quando le statistiche, i dati demografici, i rappresentanti e le caratteristiche degli utenti sono diversi nell'utenza della piattaforma rispetto alla popolazione target.[2] Quindi, se un sottogruppo demografico utilizza una determinata piattaforma, quei dati potrebbero risultare non sufficienti o inadeguati per la popolazione target.

Self-Selection Bias

Questo bias è molto simile al sampling bias, dove i soggetti della ricerca selezionano se stessi.[2] Un esempio potrebbe essere un sondaggio politico dove solo i sostenitori più entusiasti di un candidato completano il sondaggio, distorcendo così i risultati.

Social Bias

Incorriamo nel social bias quando le azioni altrui influiscono i nostri giudizi.[2] Può incorrere ad esempio con i moderni sistemi di recensione dei prodotti, dove anche un prodotto scadente può risultare appetibile se valutato positivamente dalla maggior parte dell'utenza.

Behavioral Bias

I behavioral bias derivano da diversi comportamenti degli utenti tra piattaforme, contesti o set di dati diversi.[2] Ad esempio, l'uso di emoji può variare tra le varie piattaforme, e ciò potrebbe portare a errori di comunicazione o a interpretazioni errate.

Temporal Bias

Il temporal bias deriva dalle differenze nelle popolazioni e nei comportamenti nel tempo.[2] Per esempio, un hashtag su Twitter potrebbe acquisire significati diversi nel corso del tempo, influenzando l'analisi dei trend.

Content Production Bias

Questo tipo di bias deriva dalle differenze strutturali, lessicali, semantiche e sintattiche presenti nei contenuti user-generated.[2] Le differenze di genere o culturali possono influenzare il modo in cui il linguaggio viene utilizzato, il che può a sua volta influenzare i modelli addestrati su questi dati.

2.3 Sintomi

I sintomi di fairness sono metriche o indicatori utilizzati per rilevare e misurare la presenza di bias nei dati. Questi sintomi aiutano a identificare disparità nel trattamento o nei risultati tra diversi gruppi, in particolare quelli definiti da attributi protetti come razza, genere o età. Misurare i sintomi di fairness è fondamentale per valutare l'equità di un sistema prima della fase di addestramento del modello, permettendo di intervenire preventivamente in modo da promuovere decisioni più giuste ed eque, senza la necessità di costruire e addestrare il modello.

2.3.1 Lista dei sintomi

Successivamente sono elencati i sintomi di fairness utilizzati per gli esperimenti effettuati in questa ricerca.

Kendall rank correlation coefficient

A livello matematico, il Kendall Rank Correlation Coefficient è un metodo statistico non parametrico per andare a misurare la correlazione ordinale tra due quantità misurabili. Per confrontare due insiemi ordinati definiti sullo stesso insieme, l'approccio di Kendall prevede di andare a contare il numero di coppie diverse tra questi due insiemi ordinati. Questo numero andrà ad indicare la distanza tra i due insiemi, ed è definito come symmetric difference distance. La formula per calcolare il Kendall rank correlation coefficient è la seguente:

$$\tau = 1 - \frac{2 \times [d_\Delta(P_1, P_2)]}{N(N-1)}$$

dove $d_\Delta(P_1, P_2)$ indica la symmetric difference distance tra due set di coppie ordinate P_1 e P_2. Il valore risultante τ, una volta normalizzato, varierà tra -1 e 1. Un valore equivalente a -1 corrisponde alla più grande distanza possibile tra le due variabili (ottenuto quando un ordine è l'esatto opposto di un altro ordine) mentre 1 indica la distanza più piccola possibile (ottenuto quando i due ordini sono uguali).[5]

Mutual information

La mutual information serve per andare a misurare la quantità di informazione inerente ad una variabile che è possibile ricavare da una variabile casuale. Può essere interpretata come la riduzione dell'incertezza su una variabile casuale data la conoscenza di un'altra. Più alto risulta essere il valore, minore sarà l'incertezza. In caso di valore pari a 0, le due variabili risultano essere indipendenti. Per due variabili discrete X e Y con joint probability distribution (la probabilità che due eventi si verifichino contemporaneamente) $P_{XY}(x, y)$, la mutual information tra di loro, denotata come $I(X;Y)$ è data dalla seguente formula:

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y)\ log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}}\ log\frac{P_{XY}}{P_X P_Y}$$

dove $P_X(x)$ e $P_Y(y)$ sono le probabilità marginali, mentre E_P indica il valore atteso sulla distribuzione P. Possiamo quindi definire la mutual information come la "riduzione dell'incertezza" sulla variabile X, o come "la riduzione attesa del numero di domande sì/no necessarie per indovinare X dopo aver osservato Y".[6]

Unprivileged Positive Probability (UPP) & Privileged Positive Probability (PPP)

Questi sintomi vengono utilizzati per misurare la disparità nelle probabilità di eventi positivi tra gruppi privilegiati e non privilegiati. Si osserva quindi se ci sono differenze nel modo in cui i risultati positivi vengono distribuiti tra i gruppi privilegiati e non, utilizzando le "ground truth labels" (si riferisce alle etichette o ai valori reali, effettivi e verificati di un dataset, che rappresentano la verità rispetto a un particolare compito o problema). Si identificano andando a calcolare:

$$Pr(y = 1 \mid g_j) \quad \text{e} \quad Pr(y = 1 \mid g_i)$$

dove:

- $Pr(y = 1 \mid g_j)$ indica la probabilità che un evento positivo si verifichi in un gruppo non privilegiato
- $Pr(y = 1 \mid g_i)$ indica la probabilità che un evento positivo si verifichi in un gruppo privilegiato.

Absolute Probability Difference (APD)

Misura la disparità nelle probabilità tra eventi di gruppi privilegiati e non privilegiati. Come per UPP e PPP, si utilizzano le ground truth labels per andare a calcolare $Pr(y = 1 \mid g_j)$ e $Pr(y = 1 \mid g_i)$. L'APD è la differenza assoluta tra le due probabilità. Un valore elevato indica che il modello tratti i gruppi in modo non equo, favorendone uno rispetto ad un altro in termini di probabilità di ottenere un risultato positivo.

Privileged & Unprivileged Group Unbalance

Misura la dimensione osservata e la dimensione attesa tra gruppi privilegiati e non privilegiati rispetto all'etichetta positiva. Serve per andare a comprendere se un gruppo è sovra-rappresentato o sotto-rappresentato rispetto alle aspettative. Si va a calcolare il rapporto per i gruppi privilegiati e non privilegiati e si interpreta il risultato:

- 1: il gruppo è bilanciato, cioè la dimensione osservata è uguale a quella attesa
- $>$ 1: il gruppo è sovra-rappresentato (oversampled), cioè la dimensione osservata è maggiore rispetto a quella attesa
- $<$ 1: il gruppo è sotto-rappresentato (undersampled), cioè la dimensione osservata è minore rispetto a quella attesa.

Rapporti significativamente diversi da 1 indicano che i gruppi sono sovra o sotto-rappresentati, suggerendo quindi la presenza di potenziale bias. Un gruppo sovra-rappresentato potrebbe essere più favorito dal modello rispetto ad un gruppo sotto-rappresentato.[3]

Kurtosis & Skewness

Sono due metriche che descrivono la distribuzione di un dataset, aiutando a comprendere meglio le caratteristiche dei dati. Certe volte andare a considerare solo media e varianza risulta non essere sufficiente; per questo vengono introdotti questi due sintomi. La Kurtosis (Curtosi) misura la "forma" della distribuzione, dando

informazioni riguardando l'estremità dei valori nel dataset, ovvero se ci sono outlier significativi. Valori vicino a 3 indicano una Kurtosis normale, indicando quindi una distribuzione quanto più vicino alla "normalità". La skewness (Asimmetria) misura il grado di asimmetria di una distribuzione rispetto alla sua media. Skewness uguale a 0 indica una distribuzione simmetrica. In entrambi i casi di Skewness e Kurtosis, valori che differiscono da quelli standard possono essere sintomo di bias nei dati.

Gini Index, Simpson Diversity, Shannon Entropy, Imbalance Ratio

Vengono utilizzati per valutare la distribuzione, la diversità e lo sbilanciamento dei dati.

- Gini Index: comunemente utilizzato negli alberi decisionali, misura l'impurità della suddivisione dei dati, ovvero quanto le classi risultano mescolate nel dataset.
- Simpson Diversity: misura la probabilità che due elementi estratti casualmente da un set di dati appartengano alla stessa classe.
- Shannon Entropy: un'ulteriore tecnica, oltre alla Simpson Diversity, per misurare la diversità di un dataset. Tiene conto del numero di categoria e dell'uniformità della distribuzione delle classi.
- Imbalance Ratio: metrica progettata per misurare il grado di sbilanciamento di un dataset, spesso utilizzaton ei problemi di classificazione binaria, dove le classi non sono equamente distribuite.

2.4 Fairness e datasets

Come visto in precedenza, la presenza di bias all'interno dei dati può andare a influire negativamente sulle prestazioni del modello per quanto concerne il garantire dei risultati fair per un modello di machine learning. Risulta importante quindi esplorare le relazioni tra dataset e fairness, analizzando come le caratteristiche intrinseche dei dati, come la rappresentatività e la distribuzione delle variabili sensibili, possano condurre a decisioni imparziali o, al contrario, perpetuare ingiustizie. In

questa sezione verranno analizzati alcuni tra i dataset più utilizzati nel mondo reale per l'apprendimento consapevole dell'equità, i quali verranno adoperati per il nostro studio.

2.4.1 German credit

Il german credit dataset è costituito da campioni di titolari di conti bancari in Germania. Questo set di dati viene utilizzato per la previsione della valutazione del rischio, ovvero per andare a determinare se è rischioso o meno concedere credito ad una persona.[7] Consiste in un set di 1000 istanze prive di valori mancanti. Ogni istanza presenta 11 variabili categoriche, 7 variabili intere e 2 variabili binarie, oltre alla variabile target.

2.4.2 Heart Disease

L'heart disease dataset è una raccolta di dati medici provenienti da gruppi separati di pazienti, sottoposti a procedure mediche utilizzate per diagnosticare e trattare problemi cardiaci in 5 diversi centri medici.[7] Presenta un complessivo di 76 attributi, ma nello stato dell'arte ne vengono consigliati solamente 14. La variabile target è una variabile binaria che si riferisce alla presenza o meno di malattie cardiache nel paziente.

2.4.3 Student Performance

Lo student performance dataset presenta una raccolta di dati inerenti ai risultati nell'istruzione secondaria di due scuole portoghesi. Gli attributi comprendono i voti degli studenti, caratteristiche demografiche, sociali e relative alla scuola, e sono stati raccolti utilizzando pagelle e questionari scolastici.

Tabella 2.1: Attributi del german credit dataset

Attributi	Tipo	Valori	Descrizione
checking-account	Categorica	4	Stato del conto corrente esistente
duration	Numerica	[4-72]	La durata del credito (mesi)
credit-history	Categorica	5	Storia creditizia
purpose	Categorica	10	Scopo (auto, educazione, ecc.)
credit-amount	Numerica	[250-18,424]	Importo del credito
savings-account	Categorica	5	Conto di risparmio/obbligazioni
employment-since	Categorica	5	Impiego attuale da...
installment-rate	Numerica	[1-4]	Il tasso di rata in percentuale del reddito disponibile
personal-status-and-sex	Categorica	4	Status personale e sesso
other-debtors	Categorica	3	Altri debitori/garanti
residence-since	Numerica	[1-4]	Residenza attuale da...
property	Categorica	4	Proprietà
age	Numerica	[19-75]	Età
other-installment	Categorica	3	Altri piani di rateizzazione
housing	Categorica	3	Alloggiamento
existing-credits	Numerica	[1-4]	Numero di crediti esistenti presso questa banca
job	Categorica	4	Lavoro

number-people-provide-maintenance-for	Numerica	[1-2]	Numero di persone a carico
telephone	Binaria	[Yes, None]	Numero di telefono
foreign-worker	Binaria	[Yes, No]	L'individuo è un lavoratore straniero?
class	Binaria	[1, 2]	Variabile target (1 = Good, 2 = Bad)

Tabella 2.2: Attributi del student performance dataset

Attributi	Tipo	Valori	Descrizione
school	Binaria	[GP,MS]	Scuola dello studente
sex	Binaria	[Male, Female]	Sesso dello studente
age	Numerica	[15-22]	Età (in anni)
address	Binaria	[U,R]	Il tipo di indirizzo (U = urban, R = rural)
famsize	Binaria	[LE3,GT3]	La dimensione della famiglia (LE3 = minore o uguale di 3, GT3 = più di 3)
Pstatus	Binaria	[T,A]	Stato di convivenza dei genitori (T = vivono insieme, A = vivono separati)
Medu	Numerica	[0-4]	Educazione della madre
Fedu	Numerica	[0-4]	Educazione del padre
Mjob	Categorica	5	Lavoro della madre
Fjob	Categorica	5	Lavoro del padre

reason	Categorica	4	La motivazione per scelta della scuola
guardian	Categorica	3	Il tutore dello studente (madre, padre, altro)
traveltime	Numerica	[1-4]	Il tempo di viaggio da casa a scuola
studytime	Numerica	[1-4]	Il tempo di studio settimanale
failures	Numerica	[0-3]	Il numero di bocciature passate
schoolsup	Binaria	[Yes,No]	Esiste un supporto educativo aggiuntivo?
famsup	Binaria	[Yes,No]	Esiste un supporto educativo familiare?
paid	Binaria	[Yes,No]	Sono previste lezioni extra a pagamento all'interno della materia del corso (matematica o portoghese)?
activities	Binaria	[Yes,No]	Sono previste attività extracurriculari?
nursery	Binaria	[Yes,No]	Lo studente ha frequentato la scuola dell'infanzia?
higher	Binaria	[Yes,No]	Lo studente vuole ottenere un'istruzione superiore?
internet	Binaria	[Yes,No]	Lo studente ha accesso a internet a casa?
romantic	Binaria	[Yes,No]	Lo studente si trova in una relazione sentimentale?
famrel	Numerica	[1-5]	La qualità delle relazioni familiari (1 = very bad, 5 = excellent)

freetime	Numerica	[1-5]	Tempo libero dopo la scuola (1 = very low, 5 = very high)
goout	Numerica	[1-5]	Quanto spesso lo studente esce con gli amici (1 = very low, 5 = very high)
Dalc	Numerica	[1-5]	Consumo di alcol durante la giornata lavorativa (1 = very low, 5 = very high)
Walc	Numerica	[1-5]	Consumo di alcol durante il weekend (1 = very low, 5 = very high)
health	Numerica	[1-5]	Lo stato di salute corrente (1 = very low, 5 = very high)
absences	Numerica	[0-32]	Numero di assenze scolastiche
G1	Numerica	[0-19]	Valutazione del primo periodo
G2	Numerica	[0-19]	Valutazione del secondo periodo
G3	Numerica	[0-19]	Valutazione finale

Tabella 2.3: Attributi dell'heart disease dataset

Attributi	Tipo	Valori	Descrizione
age	Numerica	[29-77]	Età
sex	Binaria	[Male, Female]	Sesso (1 = male, 0 = female)
cp	Categorica	4	Tipo di dolore al petto
trestbps	Numerica	[94-200]	Pressione sanguigna a riposo (in mm Hg)
chol	Numerica	[126-564]	Livello di colesterolo sierico
fbs	Binaria	[0,1]	Glicemia a digiuno > 120 mg/dl
restecg	Categorica	3	ECG a riposo
thalach	Numerica	[71-4]	Risultati dell'elettrocardiogramma a riposo
exang	Binaria	[0,1]	Angina indotta dall'esercizio
oldpeak	Numerica	[0-6.2]	Depressione del tratto ST indotta dall'esercizio fisico rispetto al riposo
slope	Categorica	3	Pendenza del tratto ST dell'ECG durante il picco dell'esercizio
ca	Categorica	5	Numero di vasi principali colorati dalla fluoroscopia
thal	Categoria	4	Tipo di talassemia
target	Binaria	[0,1]	Presenza o assenza di malattia cardiaca

CAPITOLO 3

Metodi di ricerca

L'obiettivo di questo studio è analizzare e valutare come i Large Language Models (LLM) operino nell'ambito della fairness nel machine learning. In particolare, si prenderanno in esame diversi dataset, sui quali verranno analizzati i principali sintomi di fairness, per comprendere come l'LLM riesca a interpretarli e valutarli. Successivamente, verranno esaminati i metodi di refactoring suggeriti dall'LLM per mitigare il bias evidenziato dai vari sintomi.

3.1 Obiettivi e quesiti di ricerca

Il presente lavoro si propone di valutare l'efficacia di un LLM nell'ambito della fairness nel machine learning. L'obiettivo finale della ricerca è quindi il seguente:

◎ **Our Goal.** Valutare se un LLM possa essere considerato uno strumento di supporto valido per effettuare refactoring su dataset, al fine di migliorare i valori relativi ai sintomi di fairness, che segnalano potenziali bias nei dati.

Sulla base di questo obiettivo, il principale quesito di ricerca che viene posto e analizzati è il seguente:

🔍 **RQ$_1$.** *Quanto efficaci sono i Large Language Models nell'identificare e correggere i sintomi di bias presenti nei dataset?*

La metodologia sperimentale adottata in questa ricerca è stata progettata per rispondere a questo quesito di ricerca.

3.2 Strumenti utilizzati

Per la buona riuscita della ricerca, la fase iniziale dello studio ha previsto una selezione accurata degli strumenti più adatti al raggiungimento dell'obiettivo finale. Questa fase ha incluso la scelta del Large Language Model (LLM) da utilizzare nei vari esperimenti, con particolare attenzione alla selezione dei dataset e dei sintomi di fairness da analizzare. Inoltre, per garantire la riuscita degli esperimenti con l'LLM selezionato, anche la scelta di un metodo di prompting efficace si è rivelata cruciale.

3.2.1 Large Language Model

Tra i vari LLM disponibili sul mercato, per questo studio è stato scelto GPT-4o. La selezione è stata fatta sulla base di vari fattori, tra cui la familiarità acquisita nell'utilizzo di GPT-4o e i suoi vantaggi rispetto agli altri modelli. In particolare, GPT-4o si distingue per la sua capacità avanzata nella comprensione del linguaggio naturale, fondamentale per l'analisi dei sintomi di fairness e per la proposta di strategie di refactoring. Questa caratteristica, combinata con l'ampia conoscenza pre-addestrata del modello, lo rendo uno strumento più che valido per identificare e

mitigare bias nei dataset, senza richiedere ulteriori addestramenti specifici. Un altro punto di forza di GPT-4o è la sua flessibilità nel prompting, che consente di ottenere risposte personalizzabili e precise, cruciali per l'esecuzione di esperimenti su dataset complessi. Inoltre, GPT-4o offre supporto continuo e un'efficienza computazionale elevata, garantendo ottime prestazioni e tempi di esecuzione contenuti rispetto ad altri LLM. Alla luce di questi vantaggi, GPT-4o è stato ritenuto la scelta più adeguata per il raggiungimento degli obiettivi della ricerca, rispetto ad altri modelli disponibili.

3.2.2 Datasets

Per questo studio, tra i vari dataset disponibili nello stato dell'arte e sul web, sono stati scelti Student Performance, Heart Disease e German Credit. La scelta è stata guidata da una serie di considerazioni e fattori chiave che rendono questi dataset particolarmente adatti all'analisi dei sintomi di fairness e alle successive operazioni di refactoring sui dati. In particolar modo, questi set di dati coprono una gamma di settori critici quali istruzione, salute e finanza, permettendo di testare l'efficacia di GPT-4o nell'identificare e mitigare bias in contesti con caratteristiche e sfide differenti. La scelta di dataset eterogenei garantisce inoltre una maggiore generalizzabilità dei risultati ottenuti dallo studio.

3.2.3 Sintomi di fairness

Per valutare in modo approfondito il concetto di fairness nei dataset selezionati, sono stati scelti diversi sintomi di fairness che rappresentano aspetti differenti del bias e delle disparità nei dati. La scelta di questi sintomi è stata guidata dalla loro predisposizione nel coprire una vasta varietà di punti di vista, consentendo così un'analisi più dettagliata e diversificata delle disuguaglianze. Di seguito, le ragioni dietro la scelta di ciascun sintomo:

- Kendall Rank Correlation Coefficient: questo indice è stato scelto per valutare la correlazione tra variabili ordinali, risultando inoltre utile per capire se è presente una relazione significativa tra variabili sensibili e risultati previsti.

- Mutual Information: la mutual information misura la dipendenza tra due variabili e può indicare quanto l'appartenenza a un determinato gruppo (privilegiato o non privilegiato) possa influenzare i risultati previsti.

- Unprivileged Positive Probability (UPP) & Privileged Positive Probability (PPP): scelto per valutare le probabilità positive relative ai gruppi privilegiati e non, consentendo così un confronto diretto delle disparità tra di essi. È fondamentale per quantificare se un determinato gruppo beneficia in modo sproporzionato di risultati favorevoli.

- Absolute Probability Difference (APD): L'APD è stato incluso per calcolare la differenza assoluta tra le probabilità di risultato positivo per i gruppi privilegiati e non privilegiati, fornendo così una misura immediata e intuitiva della disuguaglianza, in modo da rendere più semplice quantificare il livello di unfairness nei risultati previsti.

- Privileged & Unprivileged Group Unbalance: questo sintomo risulta fondamentale per verificare se i gruppi identificati dai vari protected attributes sono equamente rappresentati rispetto ai valori attesi per una classificazione positiva.

- Kurtosis & Skewness: curtosi e asimmetria sono state incluse per analizzare la distribuzione dei dati, fornendo indicazioni sulla presenza di outlier e distorsioni nella distribuzione dei risultati, che potrebbero favorire certi gruppi rispetto ad altri, compromettendo conseguentemente la fairness.

- Gini Index, Simpson Diversity, Shannon Entropy, Imbalance Ratio: sono stati scelti per valutare la diversità e la distribuzione equa delle classi all'interno del dataset. Questi indici forniscono misure sulla disparità e sulla diversità all'interno dei gruppi, oltre a dare un ulteriore modo per comprendere se i gruppi di datis ono squilibrati in modo critico.

3.2.4 Tecniche di prompting

La selezione di una buona tecnica di prompting è stata di fondamentale importanza per ottenere risposte accurate e soprattuto utili da parte dell'LLM, essendo essa

influente non solo nella qualità delle risposte generate, ma anche nella capacità del modello nell'affrontare problemi complessi in modo sistematico. Per questo studio, è stata scelta la tecnica della Chain of Thoughts (CoT), che incoraggia il modello a suddividere il processo di ragionamento in passaggi sequenziali, esplicitando ogni fase del pensiero logico dietro una risposta. Tramite questa tecnica di prompting adottata durante lo studio, è stato possibile incrementare le capacità del modello nell'affrontare problemi complessi, ridurre gli errori da parte di GPT-4o e ottenere una maggiore chiarezza nei ragionamenti effettuati e nelle scelte prese per la buona riuscita della ricerca.[8]

3.3 Metodologia sperimentale

La metodologia adottata in questo studio è stata progettata per rispondere ai quesiti principali posti dalla ricerca, volti a valutare la capacità dell'LLM di interpretare, calcolare e mitigare il bias attraverso sintomi di fairness. Ogni fase degli esperimenti è stata strutturata in modo tale da testare l'efficacia dell'LLM in ciascun passaggio del processo di analisi, refactoring e valutazione dei dati. L'obiettivo è stato quello di esaminare come l'LLM potesse essere uno strumento di supporto valido per migliorare l'equità nei dataset.

La prima fase degli esperimenti si è concentrata sulla capacità dell'LLM di interpretare correttamente i sintomi di fairness. Questa fase è iniziata verificando le informazioni preliminari che l'LLM possedeva riguardo ai sintomi di fairness in esame. Dopo tale verifica, è stato fornito all'LLM uno dei dataset selezionati dallo stato dell'arte, chiedendogli di calcolare i sintomi di fairness su di esso. L'LLM ha calcolato i valori, individuando le variabili sensibili e la variabile target per il calcolo, offrendo così un'indicazione preliminare del livello di bias presente nel dataset. Dopo il calcolo dei sintomi, l'LLM ha analizzato i risultati ottenuti, evidenziando eventuali disparità o anomalie che potessero indicare la presenza di bias nei dati. Questa fase ha permesso di valutare la sua capacità di comprendere e interpretare i valori generati, assicurando che le analisi fossero coerenti con il problema affrontato. Dopo aver analizzato i sintomi di fairness calcolati, l'LLM è stato incaricato di proporre tecniche di refactoring mirate alla riduzione del bias individuato. In particolare, l'LLM ha

suggerito metodi di refactoring, con particolare attenzione ai metodi sintetici, per modificare la struttura del dataset e ridurre il bias rilevato. In questa fase, è stata valutata la capacità dell'LLM di proporre tecniche pertinenti e adeguate alla situazione, sulla base dell'analisi dei risultati. Successivamente, è stato chiesto all'LLM di applicare la tecnica di refactoring ritenuta più appropriata per mitigare il bias rilevato. In alcuni casi, è stato necessario intervenire per correggere l'LLM, esplicitando la tecnica da utilizzare o indirizzandolo verso metodologie migliori rispetto a quelle proposte. Questa fase ha permesso di verificare l'efficacia e l'autonomia dell'LLM nella scelta e nell'applicazione delle strategie di refactoring, dimostrando la sua capacità di adattarsi e intervenire in modo autonomo. Al termine del processo di refactoring, l'LLM è stato incaricato di ricalcolare nuovamente i sintomi di fairness sulle variabili selezionate. Questa fase conclusiva ha permesso di confrontare i risultati iniziali e finali per determinare il grado di successo dell'intervento di refactoring. L'analisi e il confronto dei risultati hanno fornito un'indicazione chiara del miglioramento ottenuto in termini di equità, evidenziando se le tecniche applicate fossero efficaci nel ridurre il bias precedentemente rilevato. Il confronto tra i risultati iniziali e quelli finali ha rappresentato un passaggio cruciale per comprendere l'impatto delle modifiche sul dataset e confermare se le tecniche di refactoring suggerite dall'LLM abbiano effettivamente contribuito alla riduzione del bias e al miglioramento della fairness nel dataset.

CAPITOLO 4

Analisi dei risultati

In questo capitolo verranno analizzati i risultati ottenuti dagli esperimenti condotti durante la ricerca. L'obiettivo è valutare l'efficacia dell'LLM nel calcolare i sintomi di fairness, suggerire strategie di refactoring e ridurre il bias nei dataset esaminati. In particolare, verranno presentati i risultati iniziali dei sintomi di fairness calcolati, confrontati con quelli ottenuti dopo l'applicazione delle tecniche di refactoring, al fine di determinare il successo degli interventi effettuati. Il capitolo sarà suddiviso in sezioni, ognuna delle quali dedicata a un singolo sintomo di fairness. In ciascuna sezione verranno presentati e analizzati i risultati degli esperimenti relativi a quel sintomo sui diversi dataset utilizzati.

4.1 Analisi preliminari dei risultati negativi

Per quanto riguarda i sintomi di fairness Kendall Rank Correlation Coefficient e Mutual Information, gli esperimenti condotti sui vari dataset hanno prodotto risultati negativi. Nessuno dei dataset analizzati ha mostrato valori che indicassero un livello di bias significativo in base a questi sintomi. Questo ha comportato una difficoltà nell'identificazione di disparità che l'LLM potesse affrontare con operazioni di refactoring. Nonostante si sia comunque tentato di procedere con il refactoring, GPT-4o non è riuscito a gestire efficacemente le modifiche sui dati in assenza di evidenti sintomi di bias. In alcuni casi, le operazioni proposte dall'LLM non hanno avuto alcun impatto sui risultati, lasciando invariati i valori iniziali. In altri casi, i risultati post-refactoring hanno mostrato un peggioramento rispetto alla situazione di partenza, suggerendo che l'LLM fatichi a intervenire in maniera costruttiva quando non vi sono chiare evidenze di bias da mitigare. Questi risultati mettono in luce i limiti del modello nell'identificare e gestire bias sottili o non presenti, evidenziando come, per determinati sintomi di fairness, l'intervento dell'LLM possa risultare inefficace o controproducente.

4.2 Analisi dei risultati per ogni sintomo

Questa sezione mostra i risultati inerenti a ogni sintomo per tutti i dataset scelti per la ricerca.

4.2.1 Unprivileged Positive Probability (UPP) & Privileged Positive Probability (PPP)

Risultati iniziali e analisi

Nel dataset Heart Disease, l'LLM è stato incaricato di calcolare i valori di UPP e PPP per la variabile sensibile sex. Il modello è riuscito a calcolare correttamente i valori relativi alle probabilità di ottenere una classificazione positiva per i due gruppi. Per il gruppo privilegiato (donne), l'LLM ha calcolato una probabilità del 72,44% di ottenere un risultato positivo, mentre per il gruppo non privilegiato (uo-

mini), la probabilità è risultata essere del 42,08%. Questi calcoli hanno indicato una discrepanza significativa tra UPP e PPP, segnalando uno squilibrio evidente nella distribuzione dei risultati positivi tra i due gruppi. L'LLM ha inoltre interpretato correttamente questa disparità, evidenziando che la differenza tra le probabilità suggerisce un potenziale bias nei confronti del gruppo maschile, che ha una probabilità inferiore di ottenere una classificazione positiva rispetto alle donne. Nel caso del dataset Student Performance, l'LLM ha analizzato l'attributo sensibile sex, poiché mostrava le maggiori differenze tra UPP e PPP rispetto ad altre feature presenti. Anche in questo caso, l'LLM è stato in grado di calcolare correttamente i valori delle probabilità per i due gruppi. Per le donne (gruppo privilegiato), l'LLM ha calcolato una probabilità dell'86,9% di ottenere una valutazione positiva, mentre per gli uomini (gruppo non privilegiato) la probabilità calcolata è stata dell'81,2%. La differenza calcolata dall'LLM tra UPP e PPP, pari a circa il 7%, è stata interpretata correttamente come una leggera disparità tra i due gruppi. Pur non trattandosi di una differenza eccessiva, l'LLM ha correttamente rilevato che il gruppo delle donne gode di una probabilità leggermente superiore di ottenere un risultato positivo rispetto agli uomini. Questa interpretazione ha evidenziato la capacità del modello di riconoscere e analizzare anche differenze più sottili nella distribuzione dei risultati, suggerendo che, pur non estremo, il divario potrebbe comunque indicare un potenziale bias da affrontare. Infine, per il dataset German Credit, il sintomo è stato calcolato sulla variabile foreign_worker, che indica se l'individuo è un lavoratore straniero o meno. Tra le varie feature sensibili, questa è risultata quella con la maggiore discrepanza tra UPP e PPP, con UPP (per i lavoratori stranieri) pari a 0,693, mentre PPP (per i lavoratori non stranieri) pari a 0,892. Le istanze di lavoratori stranieri sono in numero molto inferiore rispetto alle istanze di lavoratori non stranieri, ma con una probabilità maggiore di essere classificati positivamente.

Strategie di refactoring suggerite e applicate

Nel dataset Heart Disease, la disparità tra uomini e donne nella probabilità di ottenere una classificazione positiva è stata affrontata inizialmente con una strategia di resampling. L'LLM ha suggerito un primo intervento di oversampling per

il gruppo non privilegiato (uomini) e di downsampling per il gruppo privilegiato (donne). Tuttavia, questo approccio ha portato a uno squilibrio eccessivo nel numero di campioni tra i due gruppi, con 1013 campioni per gli uomini e solo 312 campioni per le donne. Questo evidente predominio degli uomini ha evidenziato la necessità di una nuova strategia di resampling. Di conseguenza, l'LLM ha suggerito un resampling strategico per riequilibrare ulteriormente il dataset. In questa seconda fase, è stato ridotto moderatamente il numero di campioni maschili con esito negativo, accompagnato da un oversampling più controllato per il gruppo delle donne. Questa combinazione ha portato a un risultato più bilanciato, con 890 campioni per gli uomini e 624 per le donne. Questo nuovo approccio ha corretto lo squilibrio iniziale, riducendo la disparità numerica e migliorando la rappresentanza tra i due gruppi. Nel dataset Student Performance, l'LLM ha optato per una strategia di oversampling sintetico per il gruppo non privilegiato, ovvero i maschi. L'obiettivo era quello di riequilibrare la distribuzione dei campioni tra i due sessi, portando il numero di campioni dei maschi allo stesso livello di quello delle femmine. Questo intervento è stato efficace nel bilanciare il dataset e ha portato a un miglioramento significativo delle probabilità di ottenere una valutazione positiva. Nel caso del dataset German Credit, l'LLM ha incontrato le prime difficoltà nel consigliare strategie di refactoring utili a mitigare il bias rilevato attraverso il calcolo del sintomo. Dopo vari tentativi, è stato necessario fornire esplicitamente una tecnica di refactoring, optando per l'oversampling sintetico dei lavoratori stranieri con classificazione negativa, allo scopo di ridurre la PPP e avvicinarla quanto più possibile al valore della UPP. Questo esempio evidenzia come, con dataset di natura strutturale più complessa, l'LLM fatichi a scegliere autonomamente la strategia di refactoring più adatta per mitigare il bias indicato dal sintomo.

Risultati post-refactoring

Dopo l'applicazione del refactoring e il successivo ricalcolo delle probabilità di esito positivo, i risultati ottenuti per il dataset Heart Disease hanno mostrato un notevole miglioramento. I valori di UPP e PPP risultanti erano pari a 67,42% per gli uomini e 61,70% per le donne, evidenziando un bilanciamento più equo tra i

due gruppi. Nel dataset Student Performance, a seguito dell'oversampling sintetico applicato al gruppo non privilegiato (maschi), sia UPP che PPP sono stati ricalcolati e hanno mostrato un valore pari a 86,9% per entrambi i gruppi. Per quanto riguarda invece il dataset German Credit, dopo l'oversampling, la percentuale di lavoratori stranieri con esito positivo è scesa al 73,3%, mentre per i lavoratori non stranieri è rimasta invariata.

Confronto dei risultati

Nel dataset Heart Disease, i valori iniziali mostravano una disparità significativa tra uomini e donne, con una PPP del 72,44% per le donne e una UPP del 42,08% per gli uomini. Questa differenza evidenziava un forte squilibrio nelle probabilità di ottenere un esito positivo tra i due gruppi. Dopo l'applicazione delle strategie di refactoring suggerite dall'LLM, i valori di UPP e PPP si sono avvicinati considerevolmente, riducendo la disparità: 67,42% per gli uomini e 61,70% per le donne. Sebbene non si sia ottenuto un bilanciamento perfetto, il miglioramento è stato significativo, dimostrando che l'LLM è stato in grado di ridurre notevolmente lo squilibrio iniziale e migliorare l'equità tra i due gruppi. Nel caso del dataset Student Performance, la disparità iniziale tra uomini e donne era meno marcata rispetto al dataset Heart Disease, con una PPP dell'86,9% per le donne e una UPP dell'81,2% per gli uomini. Dopo l'intervento di oversampling sintetico per i maschi, l'LLM è riuscito a bilanciare completamente le probabilità tra i due gruppi, portando sia UPP che PPP a un valore pari all'86,9%. Questo risultato indica che l'intervento dell'LLM è stato pienamente efficace nel mitigare il bias, portando a una perfetta equità nelle probabilità di ottenere un esito positivo per entrambi i gruppi. Infine, nel dataset German Credit, la disparità tra UPP e PPP è scesa dal 19,9% al 4%, indicando che l'intervento ha avuto esito positivo anche in questo caso. Tuttavia, a differenza degli altri due esperimenti, è stato necessario supportare l'LLM nell'interpretazione accurata dei risultati e guidarlo verso una strategia di refactoring più adatta al problema in esame.

Tabella 4.1: Risultati Unprivileged Positive Probability (UPP) & Privileged Positive Probability (PPP)

	UPP-PPP pre-refactoring	UPP-PPP post-refactoring
Heart Disease	0.42 - 0.72	0.61 - 0.67
Student Performance	0.81 - 0.87	0.87 - 0.87
German Credit	0.69 - 0.89	0.69 - 0.73

4.2.2 Absolute Probability Difference (APD)

Risultati iniziali e analisi

L'Absolute Probability Difference (APD) è stato calcolato sui tre dataset per tutti i protected attributes individuati al loro interno. Successivamente, sulla base dei risultati, è stata considerata la feature che mostrava i valori più indicativi di bias. Per quanto riguarda il dataset Heart Disease, la feature sensibile che necessitava di maggiore attenzione è risultata essere la variabile sex. In questo caso, il gruppo privilegiato era costituito dalle donne, che presentavano una probabilità di esito positivo (presenza di malattia cardiaca) pari al 72,44%. Al contrario, il gruppo non privilegiato, composto dagli uomini, aveva un outcome positivo nel 42,08% dei casi. Sulla base di questi valori, l'APD iniziale tra i due gruppi risulta pari a 0,304, segnalando una differenza significativa nella probabilità di ottenere un esito positivo tra uomini e donne. Nel caso del dataset Student Performance, la variabile sensibile che presentava le maggiori discrepanze era address (area di residenza). Essendo la variabile target (G3) di tipo numerico, e basandosi sui criteri di valutazione portoghesi, è stata scelta una soglia di 10 per definire un esito positivo. L'APD tra i gruppi U (urbano) e R (rurale) per il protected attribute address era pari a 0,0995. Sebbene questo valore non sia particolarmente elevato, rappresentava la maggiore discrepanza individuata tra le feature sensibili del dataset, segnalando una leggera disparità nella probabilità di ottenere una valutazione positiva tra gli studenti provenienti da aree urbane e rurali. Infine, per il dataset German Credit, la variabile foreign_worker (indicante se il soggetto fosse un lavoratore straniero o meno) ha attirato l'attenzione in termini di APD, con un valore di 0,199. Questo risultato indica una disparità del 19,9%

nella probabilità di ottenere un esito positivo tra lavoratori stranieri e non stranieri, evidenziando una disparità simile a quella osservata negli altri dataset.

Strategie di refactoring suggerite e applicate

Sulla base dei risultati osservati e analizzati nella fase preliminare sui vari dataset, in ogni esperimento è stato chiesto all'LLM di suggerire le strategie di refactoring più appropriate per affrontare i problemi rilevati. Nonostante fosse stato esplicitamente indicato che l'obiettivo era intervenire in fase di pre-processing, i consigli forniti dall'LLM non sempre erano coerenti con le richieste. Per quanto riguarda l'APD, essendo un sintomo legato allo squilibrio nella rappresentazione degli esiti positivi tra gruppi privilegiati e non privilegiati, in tutti gli esperimenti è stata proposta la tecnica del resampling. In particolare, così come per UPP e PPP, nella maggior parte dei casi l'LLM ha suggerito l'oversampling per il gruppo non privilegiato e l'undersampling per il gruppo privilegiato, ad eccezione del dataset Student Performance. Nel caso del dataset Heart Disease, è stato adottato un resampling controllato, con oversampling sintetico per il gruppo non privilegiato, basato sulla distribuzione esistente dei dati, e un moderato undersampling per il gruppo privilegiato. Questa strategia ha permesso di ottenere i primi risultati, dopo i quali si è proseguito solo con l'oversampling per il gruppo non privilegiato fino al raggiungimento dei risultati finali. Per quanto riguarda il dataset Student Performance, inizialmente è stato tentato un resampling simile a quello utilizzato nell'esperimento su Heart Disease. Tuttavia, dopo ulteriori analisi sulla struttura del dataset e sulle tecniche disponibili, l'LLM ha suggerito una discretizzazione della variabile target in intervalli (bin), suddividendo i voti in fasce: "basso" per voti inferiori a 10, "medio" per voti da 10 a 14, e "alto" per voti maggiori o uguali a 15. Infine, nel caso del dataset German Credit, l'LLM inizialmente ha proposto tecniche non adatte alla fase di pre-processing, come l'uso di reti neurali o Adjusted Thresholding. Dopo aver chiarito la necessità di interventi in pre-processing, l'LLM ha suggerito un oversampling sintetico per aumentare il numero di lavoratori stranieri con esito positivo. Questo oversampling è stato eseguito selezionando casualmente valori sia dalle colonne categoriche che da quelle numeriche, mantenendo la coerenza con i dati originali del dataset.

Risultati post-refactoring

I primi risultati riguardano gli esperimenti condotti sul dataset Heart Disease. Nel primo test con la tecnica di refactoring proposta dall'LLM, i risultati sono stati positivi, seppur di entità minima. La percentuale di uomini con esito positivo post-refactoring è salita inizialmente al 42,48%, mentre quella delle donne con malattia cardiaca è scesa al 71,99%, portando l'APD a 0,295. Alla luce di questi risultati, si è deciso di continuare con l'approccio, applicando esclusivamente l'oversampling al gruppo non privilegiato, seguendo le indicazioni precedenti per avvicinare ulteriormente la percentuale di uomini con esito positivo a quella delle donne. Dopo questo secondo intervento, la percentuale di uomini con malattia cardiaca è aumentata al 62,04%, con un APD pari a 0,099. Per quanto riguarda il dataset Student Performance, si è ottenuto un riscontro positivo già dal primo test. Dopo la discretizzazione della variabile target in intervalli (bin), l'APD è stato ricalcolato per le categorie "medio" e "alto", con valori rispettivamente pari a 0,0639 e 0,0356. Infine, per il dataset German Credit, i risultati sono stati più limitati, con un APD finale di 0,162. L'LLM non è stato in grado di eseguire ulteriori iterazioni di refactoring, nonostante fossero possibili. Questo è probabilmente dovuto alla complessa struttura del dataset German Credit, in cui l'interpretazione delle variabili da parte dell'LLM richiede il supporto di un documento aggiuntivo.

Confronto dei risultati

Gli esperimenti si sono distinti sin dalla fase di analisi dei risultati pre-refactoring. Il dataset Heart Disease ha indubbiamente mostrato i segni più significativi di bias, come evidenziato dal calcolo dell'APD, e gli interventi di refactoring proposti dall'LLM si sono rivelati molto efficaci nel migliorare i valori del sintomo, garantendo una maggiore fairness. Partendo da un APD di 0,304, si è arrivati a 0,099, dimostrando un notevole impatto positivo. Nel caso del dataset Student Performance, le feature sensibili mostravano inizialmente valori di APD piuttosto contenuti, segnalando un bias limitato. Nonostante ciò, l'LLM ha operato in modo efficace anche su valori relativamente bassi, mitigando ulteriormente il leggero bias rilevato. Per quanto riguarda il dataset German Credit, i valori iniziali dell'APD erano migliori rispetto a quelli

del Heart Disease, ma richiedevano comunque attenzione. Tuttavia, a differenza del primo dataset, i risultati non sono stati altrettanto soddisfacenti, evidenziando una mitigazione molto limitata del bias. L'intervento di refactoring, in questo caso, non ha avuto un grande impatto, ma ha comunque dimostrato una discreta capacità dell'LLM di supportare il processo di fairness nel contesto del machine learning. Dai risultati emerge che, in termini di APD, l'LLM è in grado di agire efficacemente su dataset con una struttura più semplice, mentre mostra maggiori difficoltà con dataset di natura più complessa.

Tabella 4.2: Risultati Absolute Probability Difference

	APD pre-refactoring	**APD post-refactoring**
Heart Disease	0.304	0.099
Student Performance	0.0995	0.0639 "medio" 0.0356 "alto"
German Credit	0.199	0.162

4.2.3 Privileged & Unprivileged Group Unbalance

Risultati iniziali e analisi

L'esperimento relativo al Privileged and Unprivileged Group Unbalance è stato condotto calcolando la dimensione osservata e quella attesa in termini di esito positivo per i gruppi privilegiati e non privilegiati, identificati in base alle variabili sensibili presenti nei tre dataset. Nel dataset Heart Disease, il calcolo delle dimensioni osservate e attese per i due gruppi definiti dalla variabile sensibile sex ha evidenziato uno squilibrio significativo. La dimensione osservata per gli uomini era pari a 300, mentre quella attesa era 365. Per le donne, la dimensione osservata era di 226 contro una dimensione attesa di 160. I risultati hanno mostrato un group unbalance ratio pari a 0.82 per gli uomini (gruppo privilegiato), indicando che erano sotto-rappresentati rispetto a quanto atteso. Al contrario, per le donne (gruppo non privilegiato), il group unbalance ratio era di 1.41, segnalando una sovra-rappresentazione rispetto alle

attese. Questi risultati suggeriscono un chiaro squilibrio nella distribuzione degli esiti positivi tra i due gruppi, con le donne significativamente sovra-rappresentate e gli uomini sotto-rappresentati, il che evidenzia la presenza di potenziale bias nel dataset. Per quanto riguarda invece il dataset Student Performance, il bilanciamento tra i gruppi privilegiati e non privilegiati era complessivamente buono, con un leggero squilibrio riscontrato nella variabile sensibile sex. La dimensione osservata per il gruppo privilegiato (donne) era pari a 324, mentre la dimensione attesa era 316, con un rapporto osservato/atteso di 1.02. Questo valore indica che le donne erano lievemente sovra-rappresentate rispetto a quanto atteso, considerando la distribuzione delle classi e degli esiti positivi. D'altro canto, per gli uomini, il gruppo non privilegiato, la dimensione osservata era di 225, mentre quella attesa era 232, con un rapporto osservato/atteso di 0.97. Questo valore indica che gli uomini erano leggermente sotto-rappresentati rispetto alle aspettative. Nonostante lo squilibrio sia contenuto, i risultati suggeriscono che un'ulteriore analisi o intervento potrebbe essere necessario per migliorare l'equità tra i due gruppi. Gli ultimi risultati riguardano il dataset German Credit. l'analisi della variabile sensibile foreign_worker ha rivelato uno squilibrio meno marcato rispetto agli altri dataset. I ratio di rappresentazione osservata/attesa erano pari a 1.27 per i lavoratori non stranieri e 0.99 per i lavoratori stranieri. Questo indica che i lavoratori stranieri avevano una dimensione osservata quasi identica a quella attesa, il che riflette la loro scarsa rappresentanza nel dataset. Al contrario, i lavoratori non stranieri risultavano sovra-rappresentati, con un esito positivo maggiore di quanto previsto in base alla distribuzione complessiva. Questi risultati suggeriscono che, sebbene il gruppo dei lavoratori stranieri sia quasi bilanciato rispetto alle attese, i lavoratori non stranieri mostrano una significativa sovra-rappresentazione, il che potrebbe indicare una disparità nella distribuzione degli esiti positivi tra i due gruppi. In tutti e tre i casi, l'LLM è stato in grado di interpretare correttamente l'utilizzo del sintomo e come calcolarlo, oltre a effettuare una corretta analisi del risultati ottenuti sui vari dataset.

Strategie di refactoring suggerite e applicate

In tutti e tre gli esperimenti, l'LLM ha suggerito una strategia di refactoring basata sul resampling, applicando oversampling per i gruppi identificati come sotto-rappresentati e undersampling per quelli sovra-rappresentati rispetto all'etichetta positiva. Questo approccio, tuttavia, non elimina completamente il bias, poiché, una volta effettuato il refactoring del dataset, il numero di campioni e la distribuzione rispetto all'etichetta positiva vengono modificati, alterando il numero atteso di elementi appartenenti a ciascun gruppo con esito positivo. Nonostante ciò, la tecnica riesce a migliorare i valori di group unbalance, mitigando leggermente il bias. Su dataset con un numero moderato di campioni, è possibile effettuare più iterazioni fino a raggiungere ratio vicini all'1, ma non sempre questo porta ai risultati desiderati. Inoltre, questo approccio non è consigliabile per dataset di grandi dimensioni, poiché il numero di iterazioni potrebbe diventare eccessivo, generando ulteriori squilibri e bias nella rappresentazione dei campioni.

Risultati post-refactoring

Nel dataset Heart Disease, una volta effettuato il refactoring, se si considerano le dimensioni attese calcolate sul dataset originale, i ratio per il gruppo privilegiato e non privilegiato risultano entrambi pari a 1, segnalando un perfetto bilanciamento tra dimensione osservata e attesa. Tuttavia, considerando le dimensioni attese ricalcolate sul dataset sintetico, il ratio per gli uomini è passato da 0,82 a 1,28, indicando che il gruppo è passato da essere sotto-rappresentato a essere sovra-rappresentato. Una situazione analoga si è verificata per le donne, il cui ratio è sceso da 1,41 a 0,67, segnalando che ora risultano sotto-rappresentate rispetto ai valori attesi. Per quanto riguarda il dataset Student Performance, la strategia di refactoring suggerita dall'LLM non ha prodotto risultati significativi. Nonostante diversi tentativi, i valori di group unbalance erano già molto contenuti, e pertanto non sono stati osservati miglioramenti rilevanti. Infine, nel dataset German Credit, è stato applicato un undersampling moderato per i lavoratori non stranieri, inizialmente sovra-rappresentati con un ratio pari a 1,27. Dopo due iterazioni di undersampling, il valore è sceso

da 1,27 a 1,22. Tuttavia, l'LLM ha incontrato difficoltà nel proseguire con questo approccio, nonostante fosse ancora applicabile.

Confronto dei risultati

Inizialmente, nel dataset Heart Disease, il gruppo degli uomini era sotto-rappresentato con un ratio di 0,82, mentre il gruppo delle donne era sovra-rappresentato con un ratio di 1,41. Dopo il refactoring, il ratio degli uomini è aumentato a 1,28, segnalando una sovra-rappresentazione, mentre quello delle donne è sceso a 0,67, indicando una sotto-rappresentazione. Sebbene il refactoring abbia portato a un cambiamento, ha invertito lo squilibrio piuttosto che bilanciare i gruppi, introducendo una nuova forma di bias, ma comunque mostrando un miglioramento generale in termini di riduzione dello squilibrio. Nel dataset Student Performance, i valori iniziali mostravano un lieve squilibrio, con un ratio di 1,02 per le donne e 0,97 per gli uomini. Nonostante i tentativi di refactoring, i risultati finali non hanno evidenziato miglioramenti significativi. I valori sono rimasti pressoché invariati, suggerendo che, vista la situazione già bilanciata, il refactoring non ha avuto un impatto rilevante. Per quanto riguarda il dataset German Credit, i lavoratori non stranieri erano inizialmente sovra-rappresentati con un ratio di 1,27, mentre i lavoratori stranieri presentavano un ratio quasi bilanciato di 0,99. Dopo due iterazioni di undersampling, il ratio dei lavoratori non stranieri è sceso lievemente a 1,22, indicando un miglioramento, sebbene non sufficiente per raggiungere un perfetto bilanciamento. L'LLM ha incontrato difficoltà nel proseguire con ulteriori iterazioni, limitando il potenziale di miglioramento. Questi risultati evidenziano come l'LLM abbia riscontrato difficoltà significative nel migliorare i valori di questo sintomo, non riuscendo a tenere conto delle modifiche nelle dimensioni attese causate dagli interventi di refactoring. Tuttavia, in presenza di squilibri più marcati, l'LLM è stato in grado di ottenere miglioramenti, anche se minimi.

Tabella 4.3: Risultati Privileged & Unprivileged Group Unbalance

	Group Unbalance pre-refactoring	Group Unbalance post-refactoring
Heart Disease	0.82 - 1.41	1.28 - 0.67
Student Performance	0.97 - 1.02	0.97 - 1.02
German Credit	1.27	1.22

4.2.4 Kurtosis & Skewness

Risultati iniziali e analisi

A differenza degli altri esperimenti, per il dataset Heart Disease i sintomi Kurtosis e Skewness sono stati calcolati su due feature: sex, una variabile sensibile, e chol (livello di colesterolo sierico). Per i dataset Student Performance e German Credit, invece, l'analisi è stata effettuata direttamente sulla variabile target. Nel dataset Heart Disease, la variabile sex ha mostrato una distribuzione asimmetrica verso sinistra, con valori di kurtosis pari a -1,27 e skewness pari a -0,85, indicando una disparità nella rappresentazione dei sessi, con una maggiore concentrazione di uomini. Al contrario, la variabile chol ha evidenziato una distribuzione sbilanciata verso valori più elevati, con una kurtosis di 3,99 e una skewness di 1,07. Questo indica una forte presenza di outlier e una predominanza di valori di colesterolo più alti, segnalando un possibile bias legato ai gruppi con livelli di colesterolo estremi. Nel caso del dataset Student Performance, i sintomi sono stati calcolati sulla variabile target G3. I valori di skewness (-0,91) e kurtosis (2,68) hanno indicato una leggera asimmetria verso sinistra e la presenza di outlier, con punteggi estremamente bassi o alti rispetto a una distribuzione normale. Questi risultati suggeriscono che esistono alcune irregolarità nella distribuzione dei voti, sebbene non evidenzino direttamente un bias di rappresentazione tra gruppi specifici. Infine, per il dataset German Credit, la variabile binaria customer_classification ha presentato una kurtosis negativa (-1,24), segnalando una distribuzione piatta con meno outlier rispetto a una distribuzione normale. Tuttavia, essendo una variabile binaria, questa leggera asimmetria non ha un impatto significativo in termini di bias. La skewness positiva (0,87) indica una

leggera asimmetria verso destra, con una maggiore concentrazione di valori nelle classi inferiori.

Strategie di refactoring suggerite e applicate

Nel caso del dataset Heart Disease, per la variabile sensibile sex, l'LLM ha proposto di applicare un oversampling sintetico al gruppo non privilegiato, ovvero le donne, con l'obiettivo di bilanciare la distribuzione tra i sessi. Questo oversampling è stato eseguito duplicando campioni già esistenti e aggiungendo rumore ai dati, basato sulla distribuzione originale, per evitare il rischio di overfitting e mantenere una certa coerenza nella distribuzione. Questa tecnica ha permesso di ridurre la disparità nella rappresentazione dei sessi. Per quanto riguarda la variabile chol (livello di colesterolo sierico), che presentava una forte asimmetria e una significativa presenza di outlier, è stata adottata una trasformazione logaritmica. Questo intervento ha ridotto l'impatto dei valori estremi, rendendo la distribuzione più simmetrica e attenuando il peso degli outlier. Nel dataset Student Performance, l'LLM ha proposto e testato diverse tecniche per migliorare la distribuzione della variabile target G3, che mostrava una certa asimmetria. Le tecniche suggerite includevano la trasformazione logaritmica, il clipping degli outlier, la trasformazione quantile e la winsorizzazione. Dopo aver valutato l'efficacia di ciascun approccio, è stato deciso di adottare la winsorizzazione. Questa tecnica si è dimostrata la più efficace nel migliorare la simmetria della distribuzione e ridurre l'influenza degli outlier, garantendo una distribuzione più bilanciata e meno distorta. Per il dataset German Credit, invece, l'LLM ha proposto di intervenire sulla variabile binaria customer_classification attraverso un sottocampionamento della classe dominante (lavoratori non stranieri). Questa strategia aveva l'obiettivo di ridurre lo squilibrio tra le classi, migliorando la simmetria della distribuzione e garantendo una rappresentanza più equa tra lavoratori stranieri e non stranieri. Le strategie proposte dall'LLM sono risultate, in questo caso, coerenti con le probablematiche esposte. In questo caso, le tecniche di refactoring suggerite dall'LLM si sono dimostrate coerenti con i problemi di distribuzione affrontati in ciascun dataset. Le proposte sono state adeguate alle specifiche esigenze, come l'oversampling sintetico per riequilibrare la distribuzione dei gruppi sottorappresentati, o le trasformazioni

per ridurre asimmetrie e gestire gli outlier.

Risultati post-refactoring

In seguito agli interventi di refactoring sul dataset Heart Disease per la variabile sex, la distribuzione ha subito alcune variazioni significative. La kurtosis è passata da un valore iniziale di -1,27 a -2. Sebbene questo cambiamento possa sembrare rilevante, è importante ricordare che, trattandosi di una variabile binaria, una kurtosis così negativa non porta a particolari implicazioni negative per la distribuzione. Al contrario, la skewness, inizialmente pari a -0,85, ha raggiunto il valore di 0, indicando un perfetto bilanciamento tra i due gruppi rappresentati dai sessi. Per quanto riguarda la variabile chol (livello di colesterolo sierico), la trasformazione applicata ha portato miglioramenti notevoli. La skewness, che inizialmente era piuttosto elevata con un valore di 1,07, è stata ridotta a 0,23, segnalando una distribuzione molto più simmetrica rispetto alla condizione precedente. La kurtosis ha mostrato un miglioramento altrettanto significativo, passando da 3,99 a 0,88. Questo cambiamento indica che la presenza di outlier estremi è stata notevolmente ridotta, e la distribuzione dei valori di colesterolo è ora più simile a una distribuzione normale, con code meno pronunciate. Nel caso del dataset Student Performance, i risultati del refactoring hanno evidenziato una distribuzione più equilibrata e meno influenzata dagli outlier, in particolare per la variabile target G3. La skewness è stata ridotta a 0,24, un valore che indica una distribuzione quasi simmetrica, molto più bilanciata rispetto ai valori iniziali. Anche la kurtosis ha subito una variazione significativa, passando da 2,68 a -0,84. Questo cambiamento riflette una riduzione dell'effetto degli outlier più estremi, e conferma che la distribuzione dei dati è ora più omogenea e meno distorta da valori estremamente alti o bassi. Infine, per il dataset German Credit, l'applicazione del sottocampionamento sulla variabile binaria customer_classification ha portato a un miglioramento evidente. La skewness è stata completamente azzerata, raggiungendo il valore di 0, segnalando una distribuzione perfettamente bilanciata tra le classi. Allo stesso modo, la kurtosis è passata a -2, suggerendo che, nonostante la distribuzione sia ora più piatta rispetto a una distribuzione normale, la presenza di outlier è stata comunque gestita in modo efficace.

Confronto dei risultati

Per quanto riguarda il dataset Heart Disease, i risultati ottenuti mostrano un miglioramento evidente. Inizialmente, la variabile sex presentava una skewness di -0,85, segnalando una distribuzione asimmetrica con una maggiore rappresentanza degli uomini. Dopo l'oversampling del gruppo non privilegiato (donne), la skewness è stata azzerata, con un valore finale di 0, evidenziando un bilanciamento perfetto tra i due sessi. Anche se la kurtosis è passata da -1,27 a -2, non ha avuto particolari conseguenze data la natura binaria della variabile. Per la variabile chol, il miglioramento è stato altrettanto significativo. La skewness è diminuita da 1,07 a 0,23, indicando una distribuzione molto più simmetrica rispetto alla condizione iniziale. Anche la kurtosis, inizialmente alta (3,99), è stata ridotta a 0,88, segnalando una diminuzione nella presenza di outlier estremi. Questo indica che l'intervento di refactoring è stato efficace nel ridurre l'aspetto più problematico della distribuzione di chol, ovvero la presenza di outlier e l'eccessiva asimmetria, contribuendo così a mitigare il bias presente. Nel dataset Student Performance, i risultati post-refactoring indicano un miglioramento soprattutto nella gestione degli outlier. La skewness della variabile G3, inizialmente -0,91, è passata a 0,24, segnalando una distribuzione più simmetrica. Anche la kurtosis è diminuita significativamente, passando da 2,68 a -0,84, riducendo quindi il peso degli outlier estremamente bassi o alti. Questi cambiamenti mostrano che l'LLM è stato in grado di ridurre l'impatto degli outlier e di ottenere una distribuzione più equilibrata, anche se l'asimmetria iniziale non era particolarmente marcata. Infine, per il dataset German Credit, i risultati post-refactoring mostrano un netto miglioramento del bilanciamento della variabile customer_classification. Inizialmente, la skewness era di 0,87, suggerendo una leggera asimmetria verso destra. Dopo il sottocampionamento, la skewness è stata ridotta a 0, segnalando una perfetta simmetria tra le classi. La kurtosis è invece passata da -1,24 a -2, riflettendo una distribuzione più piatta e con meno outlier, ma che non ha avuto un impatto negativo vista la natura binaria della variabile. In termini generali, l'LLM si è rivelato uno strumento di supporto valido per affrontare il bias rilevato attraverso i sintomi di Kurtosis e Skewness. Nella maggior parte dei casi, le tecniche di refactoring proposte dall'LLM hanno migliorato la simmetria delle distribuzioni e ridotto la presenza di

outlier, portando a distribuzioni più equilibrate e vicine alla normalità. Sebbene in alcuni casi non sia stato possibile eliminare completamente le asimmetrie o raggiungere una distribuzione perfettamente normale, i risultati ottenuti hanno dimostrato che l'LLM è stato efficace nel mitigare le distorsioni più evidenti, contribuendo a migliorare l'equità nei dati analizzati.

Tabella 4.4: Risultati Kurtosis & Skewness

	Kurtosis & Skewness pre-refactoring	Kurtosis & Skewness post-refactoring
Heart Disease	sex: -1.27, -0.85 chol: 3.99, 1.07	sex: -2, 0 chol: 0.88, 0.23
Student Performance	2.68, -0.91	-0.84, 0.24
German Credit	-1.24, 0.87	-2, 0

4.2.5 Gini Index, Simpson Diversity, Shannon Entropy, Imbalance Ratio

Risultati iniziali e analisi

Nel dataset Heart Disease, la variabile sensibile sex ha mostrato i valori più significativi tra i vari protected attributes. In particolare, l'Imbalance Ratio era pari a 2,2853, suggerendo una predominanza di uno dei due gruppi di sesso rispetto all'altro. Sebbene questo squilibrio non sia estremamente elevato, merita comunque attenzione. Gli altri indici, come il Gini Index (0,4235) e il Simpson Diversity (0,4235), indicano una distribuzione non completamente equa tra i gruppi, mentre lo Shannon Diversity (0,8866) suggerisce una diversità più limitata rispetto alla distribuzione ideale. Questi valori riflettono un certo grado di squilibrio nella rappresentazione dei sessi, segnalando un potenziale bias che potrebbe influenzare le previsioni. Nel dataset Student Performance, due variabili hanno richiesto particolare attenzione: pstatus (stato dei genitori) e mjob (lavoro della madre). Per la variabile pstatus, l'Imbalance Ratio era di 7,11, indicando che una delle due categorie (presumibilmente genitori conviventi o non conviventi) era nettamente dominante rispetto all'altra. Anche il Gini Index e il Simpson Diversity erano piuttosto bassi (entrambi pari a 0,216), a ulteriore conferma della scarsa equità nella distribuzione tra le categorie,

con una diversità limitata, come indicato anche dal Shannon Diversity di 0,539. Questa forte disparità tra le categorie suggerisce la presenza di bias nella variabile pstatus, che potrebbe portare a previsioni ingiuste per i gruppi meno rappresentati. Per quanto riguarda la variabile mjob (lavoro della madre), gli indici di diversità hanno mostrato uno squilibrio altrettanto marcato. Il Gini Index di 0,737 e il Simpson Diversity dello stesso valore indicano una distribuzione molto concentrata in alcune categorie, con professioni nettamente più rappresentate rispetto ad altre. Lo Shannon Diversity, con un valore di 2,10, e l'Imbalance Ratio di 5,38 confermano uno squilibrio significativo nella rappresentazione dei gruppi, evidenziando una potenziale fonte di bias. Infine, nel dataset German Credit, la variabile sensibile foreign_worker ha mostrato i valori più critici in termini di distribuzione dei gruppi. I valori degli indici, come il Gini Index pari a 0,071, il Shannon Entropy pari a 0,228 e il Simpson Diversity uguale a 0,071 indicano una netta predominanza di una delle due classi (lavoratori non stranieri), confermando una diversità estremamente bassa tra i gruppi. L'Imbalance Ratio, pari a 26,03, è particolarmente preoccupante e indica un estremo squilibrio nella rappresentazione dei gruppi, con una classe molto più rappresentata rispetto all'altra, portando a un chiaro rischio di bias nelle previsioni basate su questa variabile. L'LLM è stato in grado di calcolare in modo accurato tutti i sintomi di Gini Index, Simpson Diversity, Shannon Entropy e Imbalance Ratio per ciascun dataset, fornendo analisi coerenti con i dati esaminati. I valori ottenuti hanno permesso di identificare correttamente gli squilibri presenti, evidenziando nuovamente la capacità dell'LLM nel calcolare i sintomi e analizzarne i risultati.

Strategie di refactoring suggerite e applicate

Nel caso del dataset Heart Disease, è stato applicato il solito approccio di resampling, in particolare un oversampling per il gruppo non privilegiato (donne) e un undersampling per il gruppo privilegiato (uomini). Questo ha permesso di riequilibrare la distribuzione tra i due gruppi, riducendo il predominio di una categoria rispetto all'altra e migliorando i valori di Imbalance Ratio e degli altri indici di diversità. Per il dataset Student Performance, l'LLM ha proposto e applicato una strategia di perturbazione controllata per riequilibrare le variabili sensibili pstatus (stato dei

genitori) e mjob (lavoro della madre). Questa tecnica ha consistito nel modificare leggermente i valori delle categorie più dominanti con una probabilità controllata, evitando la duplicazione di dati o la creazione di campioni artificiali. Per la variabile pstatus, è stato applicato un tasso di perturbazione del 50%, mentre per mjob è stato utilizzato un tasso del 40%, permettendo un riequilibrio delle frequenze tra le diverse categorie. Questo intervento ha ridotto gli squilibri tra le categorie e migliorato la distribuzione complessiva senza alterare in modo eccessivo la struttura del dataset. Nel dataset German Credit, è stato effettuato un oversampling mediante la creazione di dati sintetici. L'LLM ha generato nuovi campioni basati sulle distribuzioni esistenti delle feature, con l'obiettivo di riequilibrare le classi e ridurre il predominio di una categoria rispetto all'altra. Questa tecnica ha aumentato il numero di campioni per i gruppi meno rappresentati, migliorando così l'equità complessiva del dataset in termini di distribuzione delle classi. In tutti gli esperimenti, l'LLM è stato in grado di suggerire e applicare autonomamente le strategie di refactoring appropriate, senza bisogno di interventi esterni.

Risultati post-refactoring

Dopo l'applicazione delle tecniche di resampling, i risultati ottenuti per il dataset Heart Disease mostrano un notevole miglioramento in termini di bilanciamento tra le due classi. Il Gini Index ha raggiunto il valore di 0,5, segnalando una distribuzione uniforme tra uomini e donne. Il Shannon Diversity è pari a 1, il che rappresenta il massimo possibile per un dataset con due classi, indicando una perfetta diversità e un bilanciamento ottimale tra le categorie. Anche il Simpson Diversity, con un valore di 0,5, conferma che le probabilità di selezionare esempi appartenenti a classi diverse sono elevate, evidenziando una buona rappresentanza delle classi nel dataset. Infine, l'Imbalance Ratio pari a 1 assicura che il numero di esempi per uomini e donne è esattamente uguale, riducendo il rischio di bias legato alla distribuzione dei dati. Per il dataset Student Performance, i risultati ottenuti dopo la perturbazione controllata mostrano un miglioramento per entrambe le variabili sensibili analizzate. Per la variabile pstatus, il Gini Index è ora pari a 0,467, indicando una distribuzione più equilibrata. Il Shannon Diversity ha raggiunto un valore di 0,952, segnalando una

maggiore diversità tra le categorie. Anche il Simpson Diversity riflette un buon livello di equilibrio con un valore di 0,467. L'Imbalance Ratio, pari a 1,69, mostra che, pur non essendo completamente bilanciato, lo squilibrio tra le classi è stato ridotto. Per la variabile mjob, i risultati post-refactoring sono altrettanto incoraggianti. Il Gini Index è ora a 0,791, mentre il Shannon Diversity ha raggiunto 2,3, indicando una buona diversità tra le categorie di lavoro. Il Simpson Diversity è di 0,791 e l'Imbalance Ratio è ora pari a 1,82. Nel dataset German Credit, l'oversampling sintetico ha portato a un perfetto bilanciamento tra le classi della variabile foreign_worker. Il Gini Index e il Simpson Diversity sono entrambi pari a 0,5, confermando che le due classi sono distribuite in modo uniforme. Il Shannon Diversity ha raggiunto il valore massimo di 1, segnalando una diversità ottimale tra le classi. Anche l'Imbalance Ratio è pari a 1, il che indica che le due classi sono ora equamente rappresentate nel dataset. Tuttavia, è stato notato che il numero di campioni sintetici generati per i lavoratori stranieri è piuttosto elevato rispetto alla dimensione originaria della classe. Ciò suggerisce che potrebbe essere necessario considerare un approccio di refactoring più moderato in futuro per evitare di alterare eccessivamente la distribuzione originale.

Confronto dei risultati

Nel dataset Heart Disease, i risultati iniziali indicavano uno squilibrio significativo tra le due classi di sesso, con un Gini Index di 0,4235 e un Imbalance Ratio di 2,2853. Dopo il refactoring, i risultati sono migliorati notevolmente: il Gini Index è salito a 0,5, segnalando una distribuzione completamente uniforme, mentre l'Imbalance Ratio è ora perfettamente bilanciato a 1. Anche i valori di Shannon Diversity e Simpson Diversity sono passati da valori indicativi di squilibrio a valori ottimali, rispettivamente 1 e 0,5, evidenziando una rappresentazione equa tra uomini e donne. Nel caso del dataset Student Performance, i risultati iniziali per la variabile pstatus mostravano un significativo squilibrio, con un Imbalance Ratio di 7,11 e un Gini Index di 0,216. Dopo la perturbazione controllata, i valori sono migliorati, con l'Imbalance Ratio ridotto a 1,69 e il Gini Index portato a 0,467, indicando un miglioramento sostanziale nell'equilibrio tra le classi. Anche i valori di Shannon Diversity e Simpson Diversity sono aumentati, suggerendo una maggiore diversità

e una riduzione del predominio di una singola categoria. Per la variabile mjob, i valori iniziali riflettevano uno squilibrio altrettanto marcato, con un Gini Index di 0,737 e un Imbalance Ratio di 5,38. Dopo il refactoring, l'Imbalance Ratio è stato ridotto a 1,82, e il Gini Index a 0,791, segnalando un riequilibrio tra le categorie di lavoro. Anche in questo caso, i valori di Shannon Diversity e Simpson Diversity sono migliorati, indicando che la distribuzione delle professioni è diventata più equa rispetto alla situazione iniziale, sebbene permangano alcune disparità. Nel dataset German Credit, la variabile foreign_worker mostrava inizialmente valori di squilibrio estremi, con un Imbalance Ratio di 26,03 e un Gini Index di 0,071, che indicavano una netta predominanza di una classe. Dopo l'oversampling sintetico, i risultati finali hanno mostrato un perfetto bilanciamento tra le due classi, con l'Imbalance Ratio che è sceso a 1 e il Gini Index che ha raggiunto 0,5. Anche i valori di Shannon Diversity e Simpson Diversity sono migliorati notevolmente, raggiungendo rispettivamente 1 e 0,5, il che riflette una distribuzione equa tra le classi. Nel complesso, i risultati finali post-refactoring mostrano un miglioramento evidente rispetto alla situazione iniziale in tutti i dataset analizzati. L'LLM è stato efficace nel mitigare il bias rilevato attraverso gli indici di Gini Index, Simpson Diversity, Shannon Diversity e Imbalance Ratio, bilanciando meglio le classi e riducendo lo squilibrio che avrebbe potuto influenzare negativamente le previsioni del modello. Sebbene in alcuni casi si possa considerare un approccio più moderato per evitare un'eccessiva generazione di dati sintetici, l'intervento di refactoring ha complessivamente contribuito a garantire una maggiore equità nei dataset.

Tabella 4.5: Risultati Gini Index, Simpson Diversity, Shannon Entropy, Imbalance Ratio

	Gini, Simpson, Shannon, Imbalance Ratio pre-refactoring	**Gini, Simpson, Shannon, Imbalance Ratio post-refactoring**
Heart Disease	0.4235, 0.4235, 0.8866, 2.2853	0.5, 0.5, 1, 1
Student Performance	pstatus: 0.216, 0.216, 0.539, 7.11 mjob: 0.737, 0.737, 2.10, 5.38	pstatus: 0.467, 0.467, 0.952, 1.69 mjob: 0.791, 0.791, 2.3, 1.82
German Credit	0.071, 0.071, 0.228, 26.03	0.5, 0.5, 1, 1

4.3 Considerazioni finali

Sulla base dei risultati ottenuti e analizzati precedentemente, è possibile trarre delle conclusioni che rispondono alla principale domanda di ricerca di questo studio. L'obiettivo iniziale era verificare se e in che misura un Large Language Model (LLM) potesse fungere da strumento di supporto per identificare e mitigare il bias nei dataset utilizzati per l'addestramento di modelli di machine learning. Di seguito, viene riportata la risposta alla domanda di ricerca, che sintetizza i risultati ottenuti e fornisce una valutazione sull'efficacia dell'LLM nel compito proposto.

👉 **Answer to RQ_1.** Sulla base dei risultati, si evince che l'LLM è in grado di calcolare correttamente i sintomi di fairness, ma necessita di un contesto applicativo più esplicito per allineare le informazioni in suo possesso con quelle richieste. Per quanto riguarda il refactoring, l'LLM individua quasi sempre la tecnica più adatta per mitigare il bias, anche se in alcuni casi richiede indicazioni più precise. Gli interventi di refactoring sono spesso efficaci, specialmente su dataset meno complessi e con sintomi di bias più evidenti.

CAPITOLO 5

Conclusioni

Nelle conclusioni di questo lavoro, è possibile affermare che l'uso dei Large Language Models (LLM) per identificare e mitigare il bias nei dataset rappresenta un passo avanti significativo verso una maggiore equità nei sistemi di machine learning. Un sistema di machine learning che soffre di bias può generare decisioni discriminatorie, con ripercussioni significative a livello sociale. Per questa ragione, l'identificazione e la mitigazione del bias nei dataset rappresentano una sfida cruciale nello sviluppo di modelli equi e affidabili. In questo contesto, i Large Language Models (LLM) si stanno dimostrando uno strumento promettente per affrontare il problema del bias nei dati. La loro capacità di elaborare e interpretare grandi quantità di informazioni li rende utili non solo nell'identificazione di sintomi di unfairness, ma anche nel proporre strategie per il refactoring dei dati, migliorando così l'equità delle previsioni. L'impiego di un LLM come GPT-4 per rilevare e mitigare il bias nei dataset fornisce un supporto prezioso per rendere i modelli di machine learning più equi, riducendo le disparità tra gruppi privilegiati e non privilegiati. Questo studio si è concentrato esclusivamente sulla fase di pre-processing, con l'obiettivo di mitigare il bias nei dati prima che vengano utilizzati per l'addestramento di un modello di machine learning. Tuttavia, sviluppi futuri potrebbero includere l'analisi di altre fasi, come l'in-processing e il post-processing, per valutare in che misura

gli LLM siano in grado di ridurre il bias anche durante e dopo l'addestramento. In alcuni esperimenti, l'LLM ha mostrato la tendenza a proporre soluzioni che non si limitavano alla manipolazione dei dati, ma cercavano di intervenire in fasi più avanzate dello sviluppo del modello. Questo suggerisce che l'LLM potrebbe essere un supporto utile non solo nella fase iniziale di pre-processing, ma anche nelle fasi successive del processo di machine learning, aprendo così interessanti prospettive per future ricerche.

Bibliografia

[1] S. Majumder, J. Chakraborty, G. R. Bai, K. T. Stolee, and T. Menzies, "Fair enough: Searching for sufficient measures of fairness," 2022. [Online]. Available: https://arxiv.org/abs/2110.13029 (Citato alle pagine 5 e 6)

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2022. [Online]. Available: https://arxiv.org/abs/1908.09635 (Citato alle pagine 5, 7, 8, 9, 10, 11 e 12)

[3] M. Hort and F. Sarro, "Privileged and unprivileged groups: An empirical study on the impact of the age attribute on fairness," in *2022 IEEE/ACM International Workshop on Equitable Data Technology (FairWare)*, 2022, pp. 17–24. (Citato alle pagine 5 e 15)

[4] S. Caton and C. Haas, "Fairness in machine learning: A survey," 2020. [Online]. Available: https://arxiv.org/abs/2010.04053 (Citato alle pagine 5, 6, 7 e 8)

[5] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of measurement and statistics*, vol. 2, pp. 508–510, 2007. (Citato a pagina 13)

[6] P. E. Latham and Y. Roudi, "Mutual information," *Scholarpedia*, vol. 4, no. 1, p. 1658, 2009, revision #186917. (Citato a pagina 14)

[7] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *WIREs Data Mining*

and Knowledge Discovery, vol. 12, no. 3, Mar. 2022. [Online]. Available: http://dx.doi.org/10.1002/widm.1452 (Citato a pagina 17)

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903 (Citato a pagina 27)

Ringraziamenti

Desidero esprimere la mia più profonda gratitudine al professor Fabio Palomba, mio relatore, per il suo prezioso supporto, i consigli e l'incoraggiamento che mi ha offerto durante tutto il percorso di questa tesi. Un ringraziamento speciale va anche al dottor Gianmario Voria, per la sua assistenza e disponibilità costanti.

Un grazie immenso va prima di tutto ai miei genitori e a mia sorella. Avete sempre creduto in me, sostenendomi in ogni passo di questo percorso, sia nei momenti di gioia che in quelli di difficoltà. La vostra fiducia e il vostro affetto mi hanno dato la forza di affrontare ogni sfida, e avete condiviso ogni mio successo con un entusiasmo che mi ha sempre spinto a fare di più. Grazie per avermi incoraggiato, per aver gioito con me e per essere stati una costante fonte di ispirazione.

Un ringraziamento speciale va anche a tutta la mia famiglia, che non mi ha mai fatto mancare il proprio supporto. La vostra presenza e il vostro affetto sono stati fondamentali nel rendere possibile questo traguardo. Senza di voi, tutto questo non sarebbe stato realizzabile.

Un ringraziamento speciale va a Rosa, che nell'ultimo anno è diventata una figura fondamentale nella mia vita, sia sul piano accademico che personale. La tua presenza è stata essenziale, permettendomi di crescere e maturare come individuo. Con te ho riscoperto il piacere di appassionarmi a ciò che faccio, anche quando le circostanze erano tutt'altro che favorevoli. Spero di poter proseguire questa avventura insieme e che il tempo ci riservi altre sfide da trasformare in nuove gioie.

Un grande grazie va a Tullio, Gianfranco e Francesco, che hanno accolto una persona inizialmente sola e spaesata in ambito universitario, integrandola completamente nella loro amicizia. Grazie a voi, ho vissuto tanti momenti bellissimi, sia dentro che fuori dall'università. Senza il vostro supporto nei momenti più difficili e la vostra costante presenza nella mia vita, questo percorso sarebbe stato molto più arido e faticoso.

Poi c'è Riccardo, che da 20 anni è sempre rimasto al mio fianco, nonostante le difficoltà che la vita ci ha posto davanti. Mi è difficile ricordare un singolo litigio tra noi, nonostante i nostri caratteri così diversi eppure straordinariamente compatibili. Questa affinità ci ha permesso di mantenere il contatto anche se il tempo trascorso insieme è andato riducendosi sempre di più.

Un grandissimo grazie anche a Francesca, che ha condiviso con me i momenti più difficili ma anche i più belli. Nei periodi più bui, sei stata sempre lì, dimostrando di essere una motivatrice eccezionale. Gli eventi ci hanno portati ad allontanarci, ma ogni volta siamo riusciti a ritrovarci, rafforzando il nostro legame.

Per concludere, un enorme ringraziamento va ai miei amici che, anche senza essere stati nominati esplicitamente sono tra le persone più immportanti di questo percorso. Con la loro compagnia e il loro affetto hanno reso più leggeri i momenti più complicati di questo percorso. Grazie per avermi saputo strappare un sorriso anche quando le giornate sembravano interminabili e per essere stati una fonte di positività quando ne avevo più bisogno. La vostra vicinanza mi ha aiutato a ritrovare la serenità e ad affrontare ogni difficoltà con uno spirito più leggero.