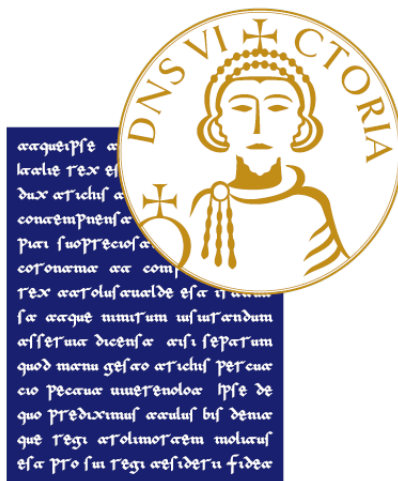


UNIVERSITÀ DEGLI STUDI DEL SANNIO
DIPARTIMENTO DI INGEGNERIA



CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

UNO STUDIO SULLE FASI DI UN
MACHINE LEARNING WORKFLOW E
UTILIZZO DI LIBRERIE

RELATORE:

CHIAR.MO PROF.

MASSIMILIANO DI PENTA

CANDIDATO:

RAFFAELE MIGNONE

399000298

CORELATORE:

DOTT.SSA

VITTORIA NARDONE

ANNO ACCADEMICO 2020-2021

[10 giugno 2021 at 18:25 -]

INDICE

1	COLLEZIONE DEI DATI E ANALISI PRELIMINARE	1
1.1	Selezione dei progetti	1
1.2	Fetch di issues e commit	1
1.3	Classificazione dei dati	2
1.3.1	Classificazione delle issues	2
1.3.2	Classificazione dei commit	4
2	ANALISI	6
2.1	RQ1: come il ML e' distribuito sull'architettura dei progetti?	6
2.2	RQ2: come sono distribuiti i bug sulle diverse fasi di ML?	8
2.3	RQ3: esiste una differenza di entropy tra ML bug e altri bug?	9
2.4	RQ4: come varia il livello di discussione tra ML bug e altri bug?	10
2.5	RQ5: come varia il time-to-fix tra ML bug e altri bug?	11
	Bibliografia	13

LISTA DELLE FIGURE

Figura 1.1	Risultati della classificazione manuale delle issues	4
Figura 1.2	Risultato della classificazione dei commit	5
Figura 2.1	Percentuale di files e directories modificate in base al tipo di cambiamento	7
Figura 2.2	Percentuale di file che utilizzano librerie di ML	7
Figura 2.3	Istanze dei fix in base alla fase	8
Figura 2.4	Entropia in base al tipo di fix	9
Figura 2.5	Livello di discussione in base al tipo	11
Figura 2.6	Giorni necessari per il fix	12

ACRONIMI

API	Application Program Interface
ML	Machine Learning
VPS	Virtual Private Server

COLLEZIONE DEI DATI E ANALISI PRELIMINARE

1.1 SELEZIONE DEI PROGETTI

L'individuazione dei progetti da analizzare è avvenuta mediante l'ausilio dell'Application Program Interface (API) messa a disposizione da GitHub. In particolare è stata eseguita una query per ottenere una lista di repository che fanno uso di librerie e framework di Machine Learning (ML) come TensorFlow, Pytorch e scikit-learn. In questo modo è stato possibile ottenere una lista di 26758 repository che è stata successivamente filtrata per individuare solo i progetti d'interesse per il seguente studio.

L'operazione di filtraggio è avvenuta attraverso due fasi; una prima automatica e una seconda manuale. La prima fase ha avuto l'obiettivo di selezionare unicamente i repository *popolari*. Nella maggior parte dei casi viene utilizzato il numero di stelle come indice della popolarità di un progetto [1], ma per questo lavoro si è preferito dare maggiore rilevanza ad altri aspetti, come il numero di fork, il numero di *contributors* e il numero di issues chiuse. Questa scelta è stata dettata dall'esigenza di selezionare non solo repository popolari, ma anche caratterizzati da una forte partecipazione della community.

I progetti che hanno superato questa prima selezione dovevano:

- essere lavori originali, per cui sono stati esclusi tutti i fork.
- avere almeno cento issues chiuse.
- avere almeno dieci contributors.
- avere almeno venticinque fork.

Alla fine di questa prima selezione il numero di repository si è ridotto a sessantasei e sono stati analizzati manualmente per rimuovere listati associati a libri e/o tutorial, progetti non in lingua inglese e librerie. Alla fine di questa seconda fase il numero di progetti è sceso a trenta.

1.2 FETCH DI ISSUES E COMMIT

Una volta individuati i progetti da analizzare si è reso necessario recuperare l'intera storia dei progetti e le issues ad essi associate. Per entrambe le operazioni è stato utilizzato il tool *perceval*[2]. Nel caso delle issues, essendo queste informazioni non direttamente contenute

all'interno del repository git, è stato necessario utilizzare nuovamente l'API di GitHub. Poiché le chiamate associate ad un singolo *token* sono limitate nel tempo si è scelto di configurare *perseval* in modo tale da introdurre in automatico un ritardo ogni qualvolta veniva raggiunto il limite. Inoltre il codice è stato dispiegato su un Virtual Private Server (VPS) in modo da poter eseguire il fetch senza che fosse necessario mantenere attiva una macchina fisica.

Con il processo precedentemente illustrato è stato possibile recuperare:

- 34180 commit.
- 15267 tra issues e pull request.

1.3 CLASSIFICAZIONE DEI DATI

1.3.1 *Classificazione delle issues*

Al fine di poter eseguire un confronto tra i *fix* di ML e quelli *generici* è stato necessario classificare sia le issues che i commit. Per quanto riguarda i primi si è scelto di attuare una classificazione basata sul testo, in particolare considerando il titolo e il corpo della issue, ma escludendo i commenti di risposta in modo da non rendere i dati troppo rumorosi. Il numero elevato di elementi non rende praticabile una classificazione manuale per cui si è optato per una classificazione automatica. A tal fine sono stati implementati ed analizzati due classificatori, uno supervisionato e uno non supervisionato.

I due modelli considerati sono:

- un classificatore statico basato su una lista di vocaboli tipici del ML.
- un modello *naïve Bayes* [3], [4].

La classificazione mediante il classificatore statico non necessita di un *labeling* manuale dei dati, ma richiede la definizione dei vocaboli tipici del ML. Lista dei termini caratteristici del Machine Learning non è stata costruita da zero, ma è basata su lavori precedenti[5]. In questo modo tutte le issues che utilizzavano almeno un vocabolo tipico del Machine Learning sono state classificate come issues di ML.

Nel caso del modello *naïve Bayes*, essendo questo un algoritmo di apprendimento supervisionato, si è resa necessaria una classificazione manuale delle issues. A tal scopo è stato eseguito un campionamento stratificato in base al progetto di provenienza di 376 issues che sono state divise tra due lettori e labelate. Durante il labeling si è scelto di classificare ulteriormente le issue di ML al fine di individuare anche

la fase in cui il problema si è palesato. La definizione delle varie fasi è avvenuta partendo da un lavoro di *Microsoft*[6].

Le fasi considerate sono:

- *Model Requirements*: questa fase comprende tutte le discussioni rispetto all'individuazione del modello più adatto, le funzionalità che questo deve esporre e come adattare un modello esistente per eseguire una diversa funzionalità.
- *Data Collection*: comprende le operazioni volte alla definizione di un dataset. Rientrano in questa fase sia la ricerca di dataset già esistenti che la costruzione di nuovi dataset.
- *Data Labeling*: questa fase si rende necessaria ogni qual volta si opera con modelli basati su apprendimento supervisionato.
- *Data cleaning*: in questa fase non rientrano soltanto le operazioni strettamente di pulizia dei dati come ad esempio rimozione di record rumorosi o incompleti, ma tutte le trasformazioni eseguite sui dati, quindi anche operazioni di standardizzazione, flip di immagini ecc.
- *Feature Engineering*: questa fase serve per identificare le trasformazioni da attuare sui dati e le migliori configurazioni degli *hyperparametri* al fine di migliorare il modello.
- *Model Training*: questa fase racchiude il training vero e proprio del modello.
- *Model Evaluation*: in questa fase vengono valutate le performance del modello utilizzando metriche standard come *precision* e *recall*, ma anche andando a confrontare i risultati ottenuti rispetto a quelli generati da altri modelli o rispetto all'esperienza¹.
- *Model Deployment*: questa fase riguarda il dispiegamento del modello sul dispositivo target.
- *Model Monitoring*: una volta dispiegato il modello deve essere continuamente monitorato al fine di assicurarsi un corretto comportamento anche sui dati reali.

A partire dal dataset *labellato* è stato possibile costruire un training e un test set, mediante i quali è stato possibile allenare e valutare le performance del modello bayesiano. Mentre le performance del primo modello sono state valutate sull'intero dataset.

Al fine di poter confrontare i due modelli sono state utilizzate le metriche di *precision* e *recall*. Com'è possibile notare dai valori riportati in *tbl. 1.1*, il modello basato sulla lista di vocaboli è leggermente più preciso del modello bayesiano, ma presenta una *recall* decisamente più bassa. Dalla *fig. 1.1a* si evince la natura minoritaria delle *issues* di

¹ Non sempre è possibile valutare un modello in modo oggettivo, ci sono determinati contesti, come ad esempio la generazione di *deep fakes*, in cui è comunque necessaria una valutazione umana per determinare la qualità del risultato.

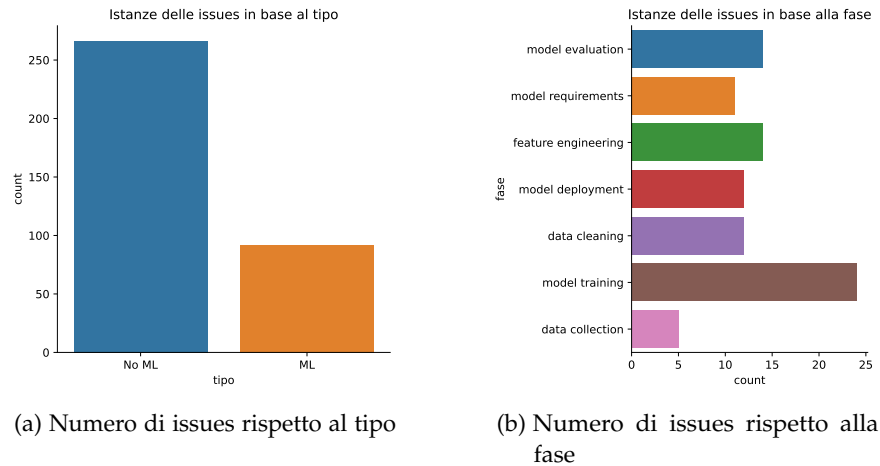


Figura 1.1: Risultati della classificazione manuale delle issues

ML rispetto alle issues generiche, per questo motivo si è scelto di preferire il modello naïve Bayes in modo da perdere quante meno istanze possibili anche a costo di sacrificare leggermente la precisione.

Tabella 1.1: Confronto dei due modelli per la classificazione delle issues.

	Classificatore statico	naïve Bayes
precision	0.46	0.41
recall	0.74	0.94

1.3.2 Classificazione dei commit

Prima di poter classificare i commit si è reso necessaria un'ulteriore fase di filtraggio in modo da poter separare i commit di *issue fixing* da quelli generici. Sono stati considerati come commit di *fix* tutti quei commit al cui interno veniva fatto riferimento a delle issues attraverso la notazione "#". Questa operazione ha ridotto il dataset dei commit a 3321 unità la cui distribuzione in base al tipo è riportata in fig. 1.2.

A questo punto è stato possibile separare i *fix* di Machine Learning da quelli generici. La classificazione è avvenuta attraverso la lista delle issues citate all'interno del *commit message* e sono stati considerati come commit di ML tutti quei commit che facevano riferimento ad almeno una issue di ML.

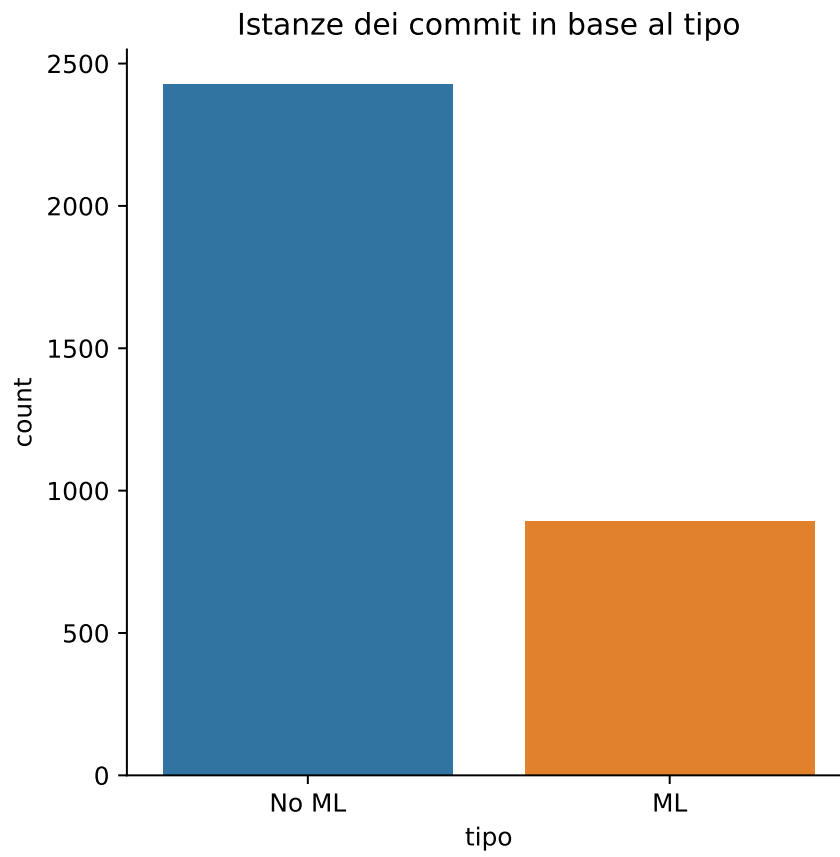


Figura 1.2: Risultato della classificazione dei commit

ANALISI

2.1 RQ1: COME IL ML E' DISTRIBUITO SULL'ARCHITETTURA DEI PROGETTI?

In questa prima analisi si è andato a verificare l'esistenza di una differenza nei files e nelle directories modificate in base al tipo di cambiamento. Per poter svolgere questa analisi è stato necessario individuare il numero totale di file modificati per *fix* generici e per i *fix* specifici del Machine Learning. A tal fine i commit sono stati raggruppati rispetto al progetto e al tipo di cambiamento (ML, no ML) e per ogni istanza di questo raggruppamento si è eseguito l'union set dei files modificati. Come output di questa fase si è generato per ogni progetto:

- l'insieme dei file modificati per *fix* di ML
- l'insieme dei file modificati per *fix* generici

Infine eseguendo l'union set tra questi due insiemi si è ottenuto l'insieme totale dei files modificati durante i *fix*. In questo modo è stato possibile andare a valutare la percentuale di files modificati in relazione al tipo di cambiamento. Attraverso la funzione di libreria Python `os.path.dirname` sono stati ottenuti i tre insiemi sopra citati anche per quanto riguarda le directories.

Dalla fig. 2.1 si può notare che i cambiamenti generici vanno ad impattare su una superficie maggiore del sistema, sia che l'analisi sia svolta al livello di files che di directories. Un'ulteriore aspetto interessante riguarda la varianza delle distribuzioni, infatti, indipendentemente dalla granularità dell'analisi, il dato riguardante i cambiamenti di Machine Learning è caratterizzato da una maggiore varianza.

Un'ulteriore analisi rispetto all'architettura dei progetti è stata svolta mediante gli import. Attraverso uno script sono stati estratti, per ogni file, gli import utilizzati all'interno del file stesso. A questo punto sono stati individuati i files di Machine Learning in base agli import utilizzati. La classificazione è avvenuta utilizzando due livelli di severità; in un primo (severità *strict*) caso sono stati considerati come import di Machine Learning solo delle librerie strettamente di ML come ad esempio `keras`, `TensorFlow`, `PyTorch`, ecc. Mentre in un secondo caso (severità *base*) sono state incluse anche librerie utilizzate

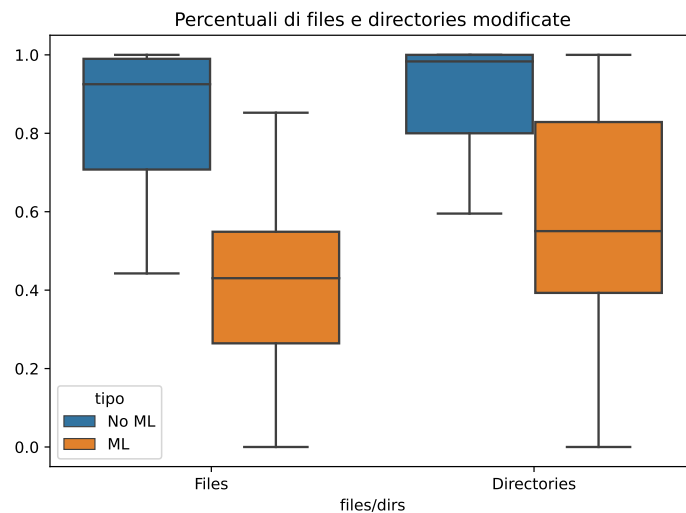


Figura 2.1: Percentuale di files e directories modificate in base al tipo di cambiamento

spesso in ambito ML, ma anche in altri ambiti, come ad esempio pandas, numpy e scipy.

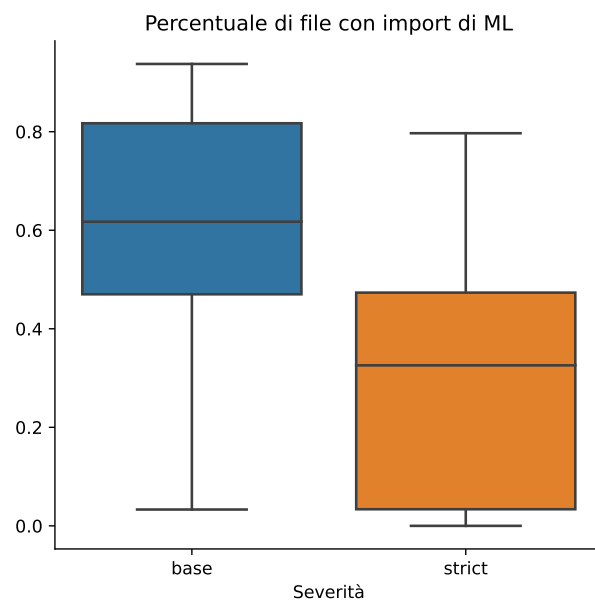


Figura 2.2: Percentuale di file che utilizzano librerie di ML

Dal boxplot riportato in fig. 2.2 si può notare che, indipendentemente dalla severità dell'analisi, la percentuale di file che utilizzano librerie di Machine Learning è caratterizzata da una forte varianza. Ciò indica che i progetti inclusi all'interno dello studio sono di varia natura e che alcuni sono più incentrati sul ML rispetto ad altri. Inoltre, considerando l'analisi *strict*, è possibile osservare come solo un 25% dei progetti abbia una percentuale di files di ML superiore al 45%.

2.2 RQ2: COME SONO DISTRIBUITI I BUG SULLE DIVERSE FASI DI ML?

Come illustrato nella sec. 1.3.2 per poter determinare la natura di un *issue fix* si è fatto ricorso alla classificazione delle *issues* ad esso associate. La maggior parte delle *issues* è stata classificata automaticamente, ma è stato comunque necessario classificarne una porzione in modo manuale per poter avere un train/test set. Come detto precedentemente, nel caso delle *issues* classificate a mano, oltre all'individuazione della tipologia (ML, non ML) è stata individuata anche la fase in cui il problema si palesava (si veda sec. 1.3.1). Questo dato aggiuntivo presente su alcune *issues* è stato *proiettato* anche sulla classificazione dei commit di *fix* per andare a valutare come questi sono distribuiti sulle varie fasi. I risultati di questa analisi sono riportati in fig. 2.3.

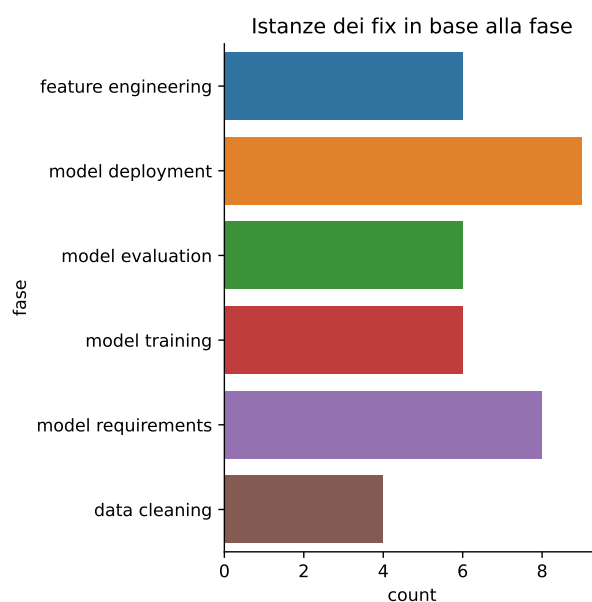


Figura 2.3: Istanze dei fix in base alla fase

Rispetto alla distribuzione sulle *issues* (fig. 1.1b) è possibile notare la scomparsa della fase *data collection*, inoltre è evidente anche la riduzione delle occorrenze di *model training* e una crescita d'importanza per quanto riguarda le fasi di *model requirements* e *model deployment*. Sfortunatamente i dati disponibili per questa analisi sono molto limitati (è stato possibile ricavare la fase solo per quaranta *fix*), per cui non è stato possibile effettuare delle analisi più approfondite.

2.3 RQ3: ESISTE UNA DIFFERENZA DI ENTROPY TRA ML BUG E ALTRI BUG?

La successiva analisi avevo lo scopo di verificare l'esistenza di una differenza tra l'entropia del *fix* rispetto alla natura di questi. L'analisi è stata svolta sia a livello di file, sia a livello di linee, quindi per ogni commit del dataset è stato necessario individuare sia il numero di file che hanno subito delle modifiche, sia il numero di linee alterate, considerando in questo modo sia le aggiunte che le rimozioni.

Inoltre per poter valutare l'entità del cambiamento è stato necessario conoscere anche il numero totale di file e di linee di ogni progetto. Questi valori sono stati calcolati attraverso la storia git del branch master¹. Per ogni commit sono stati individuati i file aggiunti (+1) e rimossi (-1) in modo tale da poter calcolare il delta-cambiamento del commit. Eseguendo la somma di questo delta su tutti i commit si è ottenuto il numero totale di file del progetto. In modo analogo si è proceduto anche per quanto riguarda le linee.

Una volta note queste informazioni preliminari è stato possibile calcolare l'entropia dei *fix* che è stata riportata nei boxplot² in fig. 2.4. Dal boxplot in fig. 2.4a è possibile notare una distribuzione equivalente per le due tipologie di *fix*. Una situazione analoga si riscontra anche nell'analisi sulle linee (fig. 2.4b) anche se in questo caso è possibile notare che i valori di entropia associati ai *fix* di ML sono shiftati leggermente verso l'alto.

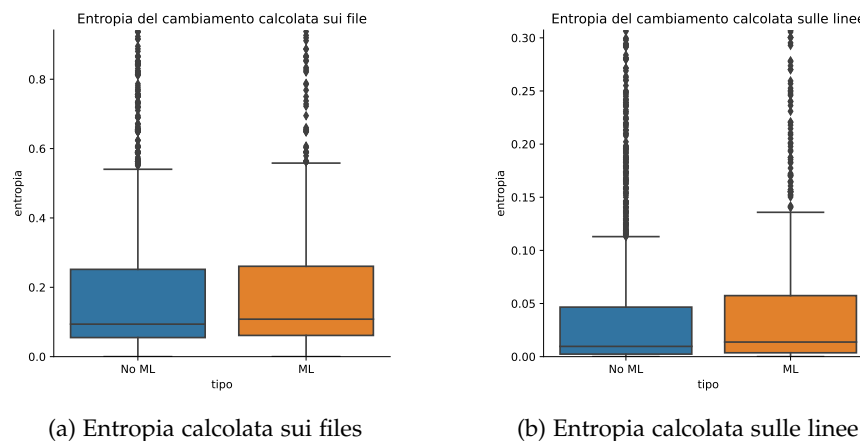


Figura 2.4: Entropia in base al tipo di fix

¹ Oltre al branch master è stato considerato anche il branch main diventato molto comune dopo le proteste del movimento Black Lives Matter e il branch master-V2 unico branch utilizzato da un progetto.

² Per ragioni di visualizzazione è stato scelto il 95-esimo quantile come limite superiore di entrambi i grafici.

Per verificare la rilevanza statistica di questa diversità sono stati svolti il *ranksum* test e il *Cliff's delta* i cui risultati sono riportati nella tbl. 2.1. Nel caso dell'entropia sui file possiamo dire che la differenza è marginale poiché il *p-value* è prossimo a 0.05, mentre nel caso dell'entropia calcolato sulle linee la differenza viene confermata dal test. In entrambi i casi, l'*effect size* è trascurabile.

Tabella 2.1: Risultati dei test statistici per quanto riguarda l'entropia

	ranksum p-values	Cliff's delta
file entropy	0.059	0.044
line entropy	5.932e-06	0.105

2.4 RQ4: COME VARIA IL LIVELLO DI DISCUSSIONE TRA ML BUG E ALTRI BUG?

Per rispondere a questa domanda è stato necessario andare a valutare il numero di commenti presenti all'interno di ogni issues. Poiché un singolo commit può far riferimento a più issues è stato considerato il numero di commenti medi. I risultati ottenuti sono stati riportati nel boxplot³ in fig. 2.5a.

In questo caso si evince una differenza molto più marcata tra le due distribuzioni. In particolare è possibile notare che le *issue fix* di ML presentano una maggiore discussione e anche una maggiore varianza. Se consideriamo la differenza interquartile, in modo da escludere completamente eventuali outlier, possiamo osservare che nei *fix* generici questa varia tra zero e uno. Ciò vuol dire che il 50% interno delle issues o non presenta commenti o ne presenta uno solo. Mentre la differenza interquartile dei *fix* di Machine Learning è compreso tra uno e cinque quindi nel 50% interno tutte le issues hanno almeno un commento di risposta.

A questo punto si è cercato di capire se al maggior numero di commenti è associata effettivamente una maggiore quantità di informazioni scambiate. Per svolgere questa analisi si è partiti dal presupposto che la quantità di informazioni scambiate sia proporzionale al numero di parole utilizzate nel commento. Quindi per ogni *issue* è stato calcolato il numero medio di parole presenti all'interno di un commento. I risultati di quest'ulteriore analisi sono riportati in fig. 2.5b. Anche in questo caso si può vedere che nel caso di ML *fix* la distribuzione presenta valori più elevati e maggiore varianza. Per cui non solo nei *fix* di Machine Learning c'è maggiore discussione, ma la discussione è anche più *densa*.

³ In questo caso il limite superiore è pari al 97-esimo quantile.

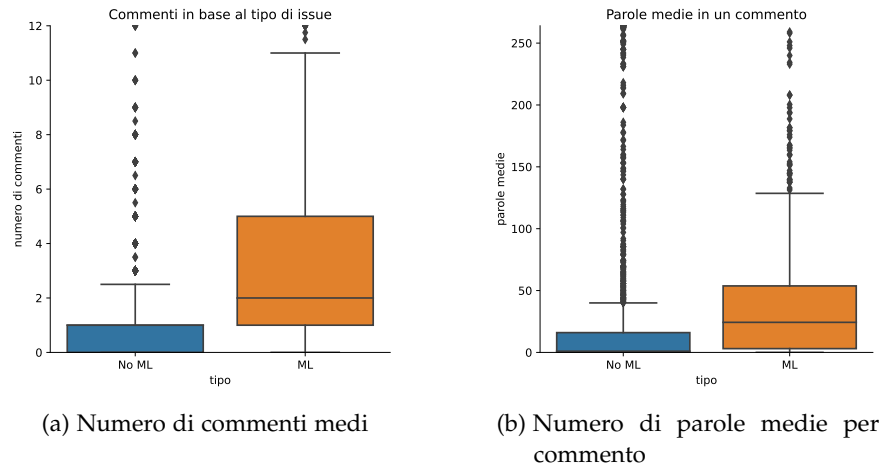


Figura 2.5: Livello di discussione in base al tipo

Anche in questo caso sono stati svolti i test statistici. In tbl. 2.2 è possibile vedere come per entrambe le metriche considerate il p -value sia abbondantemente inferiore alla soglia di 0.05 quindi abbiamo una conferma della diversità delle due distribuzioni riscontrata dal boxplot. Inoltre, per entrambe le metriche, abbiamo un *effect size* medio.

Tabella 2.2: Risultati dei test statistici per quanto riguarda il livello di discussione

	ranksum p-values	Cliff's delta
commenti medi	9.053e-75	0.425
parole per commento	2.889e-59	0.377

2.5 RQ5: COME VARIA IL TIME-TO-FIX TRA ML BUG E ALTRI BUG?

In quest'ultima analisi si vuole andare a valutare se c'è differenza nel tempo necessario per eseguire il *fix*. Per valutare questo parametro è stato necessario estrarre da ogni *issue* la data di apertura e di chiusura e calcolare i giorni che intercorrono tra queste. I risultati così ottenuti sono stati riportati in fig. 2.6.

Anche in questo caso è possibile notare una netta differenza tra i *fix* di ML e gli altri. In particolare i bug di Machine Learning necessitano, mediamente, di maggior tempo per essere risolti e sono caratterizzati da una varianza maggiore.

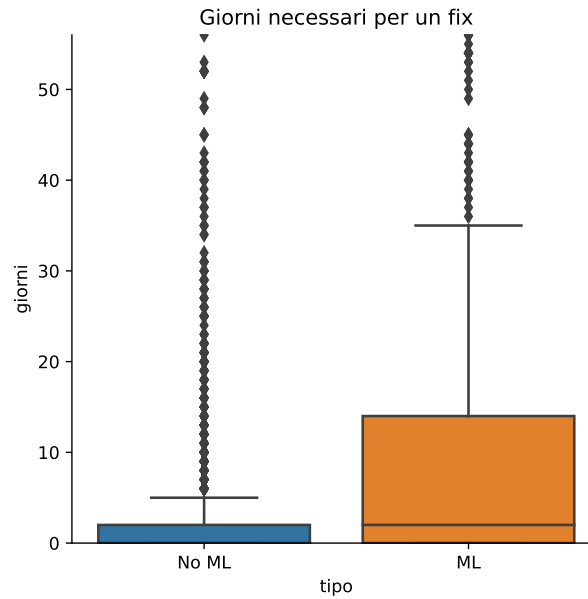


Figura 2.6: Giorni necessari per il fix

Inoltre è possibile vedere come la mediana non sia centrata, bensì spostata verso il basso. Questo vuol dire che il 50% basso dei *bug* di ML viene comunque risolto in tempi brevi (due giorni circa), mentre l'altro 50% può richiedere una quantità di tempo decisamente superiore.

Un'ulteriore testimonianza del maggior tempo necessario per risolvere le problematiche legate al ML ci viene data dagli outlier. Nel caso di un problema generico, questo, viene considerato come *anomalo* se per essere risolto necessita di un tempo superiore ai cinque giorni. Mentre nel caso dei *fix* di Machine Learning per essere considerato outlier un *issue*, necessaria di un *time-to-fix* superiore ai trentacinque giorni.

Il maggior tempo necessario ad attuare la correzione indica che i *bug* di ML sono più difficili da individuare e correggere rispetto a quelli generici. Inoltre questo risultato contribuisce a spiegare il dato emerso dalla sezione precedente, in quanto per individuare la fonte del problema sembrerebbe essere necessaria una discussione più approfondita.

Anche per quest'ultima RQ sono stati svolti i test statistici illustrati precedentemente. Dai risultati riportati in tbl. 2.3 è possibile notare un *p-value* inferiore a 0.05 e un *effect size* medio.

Tabella 2.3: Risultati dei test statistici per quanto riguarda il time-to-fix

	ranksum p-values	Cliff's delta
day-to-fix	7.354e-53	0.355

BIBLIOGRAFIA

- [1] H. Borges, A. Hora, e M. T. Valente, «Understanding the Factors That Impact the Popularity of GitHub Repositories», *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pagg. 334–344, ott. 2016, doi: 10.1109/ICSME.2016.31.
- [2] S. Dueñas, V. Cosentino, G. Robles, e J. M. Gonzalez-Barahona, «Perceval: Software Project Data at Your Will», in *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings*, mag. 2018, pagg. 1–4, doi: 10.1145/3183440.3183475.
- [3] «Naive Bayes Classifier», *Wikipedia*. mag. 21, 2021, Consultato: giu. 03, 2021. [Online]. Disponibile su: https://en.wikipedia.org/w/index.php?title=Naive_Bayes_classifier&oldid=1024247473.
- [4] P. Harrington, *Machine Learning in Action*. Manning Publications, 2012.
- [5] N. Humbatova, G. Jahangirova, G. Bavota, V. Riccio, A. Stocco, e P. Tonella, «Taxonomy of Real Faults in Deep Learning Systems», nov. 07, 2019. <http://arxiv.org/abs/1910.11015> (consultato mar. 17, 2021).
- [6] S. Amershi *et al.*, «Software Engineering for Machine Learning: A Case Study», in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, mag. 2019, pagg. 291–300, doi: 10.1109/ICSE-SEIP.2019.00042.