



UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

# INFERENZA STATISTICA NELL'ANALISI SULLA CORRELAZIONE TRA FEATURE E COVERAGE DI UN TEST CASE

RELATORE

**Prof. Fabio Palomba**

Dott.ssa **Valeria Pontillo**

Università degli studi di Salerno

CANDIDATO

**Olexiy Lysytsya**

Matricola: 0512108083

## Sommario

Il test del software è il processo di valutazione e verifica del corretto funzionamento di un'applicazione o di un prodotto software rispetto alle aspettative. I vantaggi del test includono la prevenzione dei bug, la riduzione dei costi di sviluppo e il miglioramento delle prestazioni, oltre che ad offrire al business una prospettiva oggettiva del prodotto considerato. Si stima che almeno la metà dell'intero costo di sviluppo di un progetto sia investita nel processo di testing del prodotto che si sta sviluppando[2][6]; ragion per cui l'automatizzazione del processo di testing è di forte interesse dalla comunità scientifica da almeno due decenni a questa parte. Come già riscontrato in molteplici altri ambiti, uno degli strumenti dal potenziale più significativo per l'automatizzazione del testing è costituito dall'utilizzo di algoritmi d'intelligenza artificiale adottando tecniche di machine learning[1].

Un test case è l'elemento atomico protagonista durante la fase di testing, ogni test case ha diverse caratteristiche quali la **line coverage**, una metrica descrittiva del numero di linee di codice coperte del singolo test case rispetto al source project che si sta analizzando.

Nelle seguenti pagine è descritto lo sviluppo di un modello di regressione tale che, date in input diverse caratteristiche di un test case, restituisca la predizione del numero di righe di codice ricoperte dal suddetto caso di test. **Un modello di questo tipo si basa sull'ipotesi che vi sia una correlazione tra le diverse caratteristiche di un test case ed la line coverage.** Le analisi statistiche descritte nel seguente articolo così come il training del modello sono state fatte su un dataset consistente in una raccolta di casi di test inerenti a diciotto diversi progetti software. Il frutto di questa tesi consiste quindi in un modello tale che, prese in input diverse feature statiche, restituisce in output una variabile dipendente indice della coverage associata.

<b>Indice</b>	<b>2</b>
<b>1 Introduzione</b>	<b>1</b>
1.1 Motivazioni e Obiettivi . . . . .	1
1.2 Risultati . . . . .	2
1.3 Struttura della tesi . . . . .	2
<b>2 Stato dell'arte</b>	<b>3</b>
2.1 Modello di regressione . . . . .	3
2.1.1 Regressione lineare semplice . . . . .	4
2.1.2 Cost function . . . . .	5
2.1.3 Ordinary Least Squares . . . . .	6
2.1.4 Gradient Descent . . . . .	7
2.1.5 Regularization . . . . .	7
2.2 Compromesso tra Bias e Varianza . . . . .	7
<b>3 Data analysis</b>	<b>9</b>
3.1 Descrizione delle features . . . . .	11
3.2 Analisi delle features . . . . .	13
3.3 Variabile dipendente . . . . .	13
3.4 Violin plot . . . . .	15
3.5 Indice di correlazione di Pearson . . . . .	17

<i>INDICE</i>	<b>3</b>
<b>4 Il modello</b>	<b>20</b>
4.1 Ridge model . . . . .	21
4.2 Lasso model . . . . .	22
4.3 Extreme Gradient Boosting (XGBoost) . . . . .	23
4.4 Random Forest Regression . . . . .	23
<b>5 Considerazioni e Risultati</b>	<b>25</b>
<b>Ringraziamenti</b>	<b>29</b>
<b>Bibliografia</b>	<b>30</b>

### 1.1 Motivazioni e Obiettivi

Il software testing (collaudo del software) è una delle fasi costituenti il processo di sviluppo di un prodotto software. Secondo uno studio del National Institute of Standards and Technology del 2002, si stima che nel complesso i bug nei software hanno avuto un costo pari a 59.5 miliardi di dollari l'anno per l'economia statunitense. Più di un terzo di questa spesa potrebbe essere potenzialmente evitata con l'adozione di un testing del software adeguato[11].

Lo sviluppo del software è un processo dispendioso e ad alta intensità di lavoro umano, l'interesse comune è quello di andare a semplificare tale processo in tutte le sue fasi.

L'automazione del testing contribuirebbe quindi sia alla riduzione dei tempi richiesti per lo sviluppo di un prodotto software, ma anche ad aumentare l'affidabilità del suddetto prodotto aiutando lo sviluppatore tramite strumenti di generazione automatica di casi di test applicati al codice appena scritto. Una delle metriche più interessanti che uno strumento di generazione automatica del codice può fornire allo sviluppatore è la line coverage; ovvero il quantitativo di linee di codice ricoperte dal test case che verrebbe generato rispetto al codice considerato. Avere un riferimento del line coverage stimato è un parametro che il developer può considerare per valutare l'efficacia e l'adeguatezza del test appena generato; oltre che ad essere un'informazione d'interesse al software di generazione del test stesso per il conseguimento del suo lavoro.

L'obiettivo di questa tesi è quindi costituito dall'analisi statistica sulle diverse caratteristiche (features) descrittive di un generico caso di test e la variabile dipendente che esprime il line coverage. Costruire quindi un modello di regressione in grado di andare a predire nel modo più accurato possibile il line coverage di un generico caso di test dato in input al modello.

## 1.2 Risultati

Una volta applicati metodi statistici di correlazione lineare alle diverse (feature) dei casi di test presenti nel dataset considerato, è stato possibile determinare quali siano le feature utili alla predizione del line coverage e quali invece hanno un impatto potenzialmente negativo per il training del modello.

A questo punto è stato possibile sviluppare un modello di regressione 'Random Tree Forest' in grado di predire il line coverage di un generico test case preso in input, con uno scarto assoluto medio di 1,51.

## 1.3 Struttura della tesi

Il contenuto di questa tesi è composto da tre parti, inizialmente nei prossimi capitoli vi sarà un'infarinatura generale sugli argomenti principali trattati, a seguire l'analisi del dataset e a finire lo studio del modello con annessi risultati.

### 2.1 Modello di regressione

La regressione statistica consiste nello studio della regressione verso la media. La regressione lineare è una tecnica statistica utilizzata per studiare la relazione tra due o più variabili. Da un punto di vista delle formule, la regressione lineare è una funzione matematica basata dall'equazione della retta.

Un modello di regressione lineare è composto da:

- Una sola variabile dipendente (detta anche risposta o  $Y$ )
- Una o più variabili indipendenti (dette anche  $X$  esplicative o regressori)
- Un coefficiente di regressione per ogni variabile esplicativa più un coefficiente per l'intercetta  $B$
- Un termine di errore ( $e$ ). Questo perché la relazione tra due variabili non è quasi mai perfettamente riassumibile tramite un'equazione matematica per diverse possibili ragioni: la relazione potrebbe non essere lineare; potrebbero esserci altre variabili (non considerate e/o non osservabili) che influiscono sulla  $Y$  ; ci potrebbero essere errori di misurazione delle variabili.

Nello specifico, la variabile dipendente  $Y$  è determinata dai valori dell'intercetta ( $B_0$ ) a cui vengono sommati i valori delle variabili esplicative (le  $X$ ) moltiplicate per i loro coefficienti

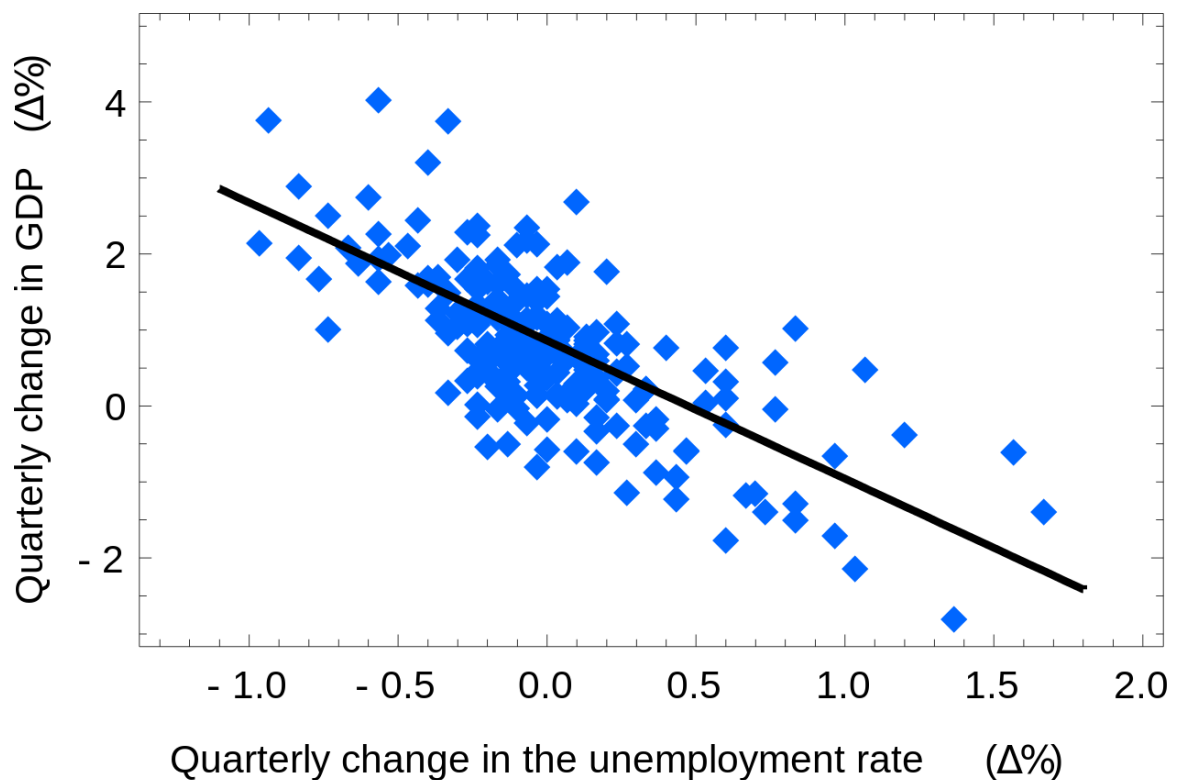
(B), più un termine d'errore (e). Con p regressori l'equazione quindi è: Figura 2.1 a cui va sommato l'errore

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_p \cdot X_p$$

Figura 2.1

### 2.1.1 Regressione lineare semplice

Qualora si abbia una singola variabile indipendente si tratta di **simple linear regression**. Riguarda punti bidimensionali le cui dimensioni sono una la variabile dipendente e l'altra la variabile indipendente, convenzionalmente associati alle coordinate x e y del piano cartesiano. L'obiettivo è trovare una funzione lineare (una linea retta non verticale) che, nel modo più accurato possibile, cerchi di predire il valore della variabile dipendente in funzione della variabile indipendente.



**Figura 2.2:** Esempio di regressione lineare semplice ove si presume una relazione lineare tra la la variabile dipendente (crescita del GDP) e cambiamenti nel tasso di disoccupazione



In genere la regressione lineare semplice prevede il calcolo di proprietà statistiche quali media, deviazione standard, correlazione e covarianza. Tale tipologia di regressione lineare difficilmente risulta utile nella pratica.

### 2.1.2 Cost function

Prima di continuare con le diverse tipologie di regressione lineare, occorre soffermarsi un attimo sul concetto di 'funzione di costo'. Nel machine learning tale funzione aiuta a determinare lo scarto tra il valore predetto dal modello rispetto al valore effettivo. Esiste poi la **loss function**, utilizzata per quantificare la 'loss' (lo scarto descritto prima) ottenuto nella fase di training. Tali funzioni sono fondamentali per algoritmi che prevedono tecniche di ottimizzazione. Ricapitolando quindi per loss function s'intende l'errore in una singola delle diverse istanze costituenti la fase di training; per cost function si fa riferimento all'average of the loss functions over an entire training dataset.

Come già detto prima, una funzione di costo ci consente di arrivare alla soluzione ottima poiché offre un criterio presso cui è possibile confrontare diverse soluzioni e determinare la migliore. Ci sono diverse tipologie di funzioni di costo ed è possibile andare a fare una prima classificazione in base alla tipologia del modello su cui si sta lavorando. Per quanto riguarda i modelli di regressione, la funzione di costo prende il nome di **Regression cost function**. Tali funzioni vengono calcolate on the distance-based error con la seguente forma:  $\text{Error} = y - y'$  dove  $y$  è l'input e  $y'$  è il valore predetto in output. Tra le regression cost function più comuni troviamo:

- **Mean Error (ME)** Questa funzione di costo prevede il calcolo dello scarto per ogni istanza del training data, per poi restituire il valore medio di essi.

Gli errori possono assumere valori negativi, ciò implica che andando a fare la media potrebbe annullarsi il valore della funzione di costo, per questo motivo tale funzione di costo viene più che altro vista come un punto di partenza per le altre

- **Mean Squared Error (MSE)** In questa tipologia di funzione di costo viene applicato il quadrato agli errori, risolvendo il problema di annullamento del risultato della funzione di costo. Una conseguenza dell'applicazione del quadrato ai singoli errori è una maggiore rilevanza attribuita anche alle piccole deviazioni, le quali vengono accentuate. Analogamente però lo stesso discorso regge anche per i cosiddetti 'outliers', i valori anomali considerevolmente più grandi rispetto alla media, assumeranno un'importanza ancora più grande. Per questo motivo si dice che MSE is less robust to outliers.

- **Root Mean Squared Error (MSE)** In questa tipologia di funzione di costo viene applicato il quadrato agli errori, risolvendo il problema di annullamento del risultato della funzione di costo. Una conseguenza dell'applicazione del quadrato ai singoli errori è una maggiore rilevanza attribuita anche alle piccole deviazioni, le quali vengono accentuate. Analogamente però lo stesso discorso regge anche per i cosiddetti 'outliers', i valori anomali considerevolmente più grandi rispetto alla media, assumeranno un'importanza ancora più grande. Per questo motivo si dice che MSE is less robust to outliers.

$$\text{MSE} = \frac{\sum_{i=0}^n (y - y')^2}{n}$$

**Figura 2.3:** MSE = (sum of squared errors)/n

- **Mean Absolute Error (MAE)** Questa funzione di costo risolve il problema di annullamento del valore restituito dal ME diversamente, anziché fare il quadrato dei singoli scarti viene invece applicato il modulo. Di conseguenza tale funzione di costo è più prestante dinanzi ai noise e outliers.

$$\text{MAE} = \frac{\sum_{i=0}^n |y - y'|}{n}$$

**Figura 2.4:** MSE = (sum of absolute errors)/n

### 2.1.3 Ordinary Least Squares

Quando si hanno più di una variabile indipendente input è possibile ricorrere alla seguente tecnica di regressione lineare. Tale procedura consiste nell'andare a minimizzare la somma dei scarti quadratici. In altre parole, data una linea di regressione sui dati, viene calcolata la distanza da ogni singola istanza di dato rispetto alla linea di regressione, calcolarne il quadrato, ripetere il procedimento per tutti le istanze dei dati e calcolarne la somma. Il valore ottenuto è quello che si cerca di minimizzare con la seguente tecnica di regressione. Dal punto di vista geometrico, tale valore corrisponde alla somma quadratica delle distanze, parallele all'asse della variabile dipendente, tra ogni istanza del dataset e il corrispondente punto sulla superficie di regressione, minore è la differenza (distanza), superiore è la capacità predittiva del modello.

### 2.1.4 Gradient Descent

Per gradient descent s'intende un algoritmo che trova iterativamente la linea di regressione più appropriata da applicare ad un dataset di training; ad ogni iterazione si va a minimizzare il margine d'errore del modello. Inizialmente al modello verranno assegnati ad ogni coefficiente dei valori casuali; viene poi calcolata la somma degli scarti quadratici per ogni coppia di valori input/output. Viene selezionato un tasso di crescita (learning rate alpha parameter) che influenzerà la "resistenza" al miglioramento tra un'iterazione e l'altra. Tale processo continuerà finché non si raggiungerà un ottimo o una situazione di stallo, ovvero tra un'iterazione e l'altra non c'è alcun miglioramento.

### 2.1.5 Regularization

La regolarizzazione è una tecnica nel machine learning che mira a generalizzare il modello, ciò implica che il modello sia performante non solo in fase di training o di testing, ma anche con un generico input che riceverà in futuro. Fondamentalmente la regolarizzazione consiste nel minimizzare the sum of the squared error del modello sul training data (applicando la tecnica dell'Ordinary Least Squares) e al tempo stesso ridurre la complessità del modello.

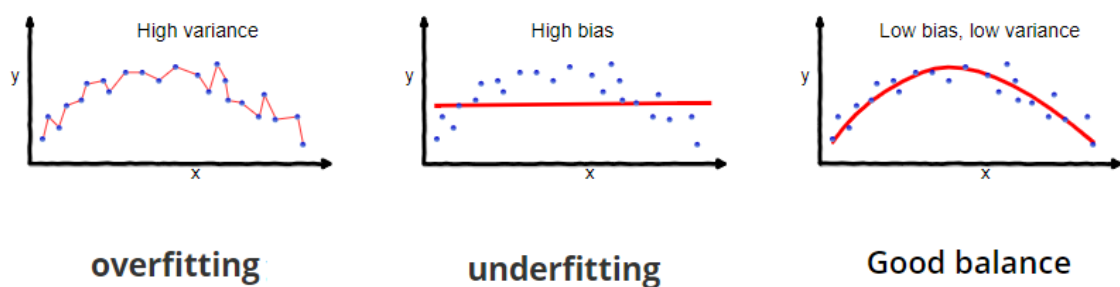
## 2.2 Compromesso tra Bias e Varianza

Per **bias** s'intende la differenza tra l'average della predizione del modello e il valore effettivo che stiamo cercando di predire. Lo si può considerare quindi come lo scarto tra il valore predetto e la realtà.

Un modello con alto bias da poco conto alle informazioni presenti nel dataset utilizzato in fase di training, il modello è troppo semplice per poter riuscire ad analizzare sufficientemente la realtà, non è in grado di considerare le variazioni, si tratta di una situazione di **underfitting**. Nel caso della **varianza** siamo di fronte ad una situazione opposta a quella appena descritta nel caso del bias; la varianza descrive la capacità del modello di considerare ed apprendere nel complesso tutte le informazioni contenute nel dataset, includendo informazioni 'scomode' come il rumore e gli outliers. Ne consegue che il modello apprende troppo dal training data a tal punto che, in fase di testing non sarà in grado di offrire predizioni accurate; ci troviamo quindi in una situazione di **overfitting**.

Nell'apprendimento supervisionato, ci si potrebbe imbattere in una situazione di under-

fitting quando il modello non riesce a concettualizzare i pattern intrinseci nel data. Questo genere di modelli avranno un alto bias ed una bassa varianza. In genere ciò accade quando non si ha un quantitativo sufficiente di data a disposizione per il training oppure ad esempio quando si sta cercando di costruire un modello di regressione lineare utilizzando però data non lineare. Una situazione di overfitting è frutto di un dataset rumoroso che confonde l'apprendimento del modello facendogli prendere in considerazione rumore o pattern fuorvianti; in genere questi modelli hanno un basso bias ed un'alta varianza.



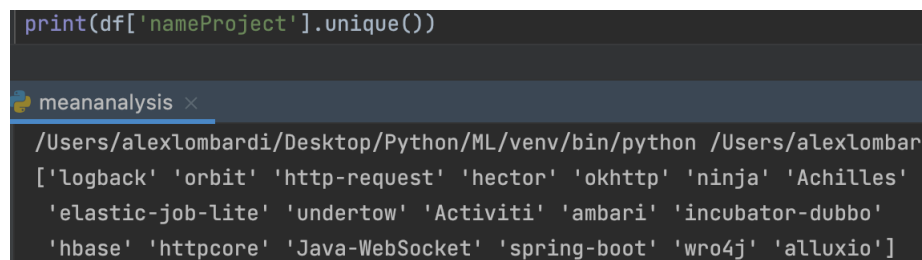
**Figura 2.5**

Il bias e la varianza sono due concetti uno l'opposto dell'altro; se il modello è troppo semplice e con pochi parametri, c'è il rischio che s'incomba in una situazione ad alto bias e bassa varianza; viceversa se il modello ha un alto numero di parametri potremmo avere un'alta varianza ed un basso bias. Occorre trovare un equilibrio prevenendo sia l'overfitting che l'underfitting sui dati. Questo compromesso nella complessità è il cosiddetto tradeoff between bias and variance: un algoritmo non può essere più complesso e meno complesso allo stesso tempo.

Il dataset descritto nel seguente capitolo è già stato protagonista di altri articoli scientifici; ad esempio uno studio sulla flakiness dei test cases, in altre parole un'analisi sulla non deterministicità riguardante l'esito dei suddetti test cases [12].

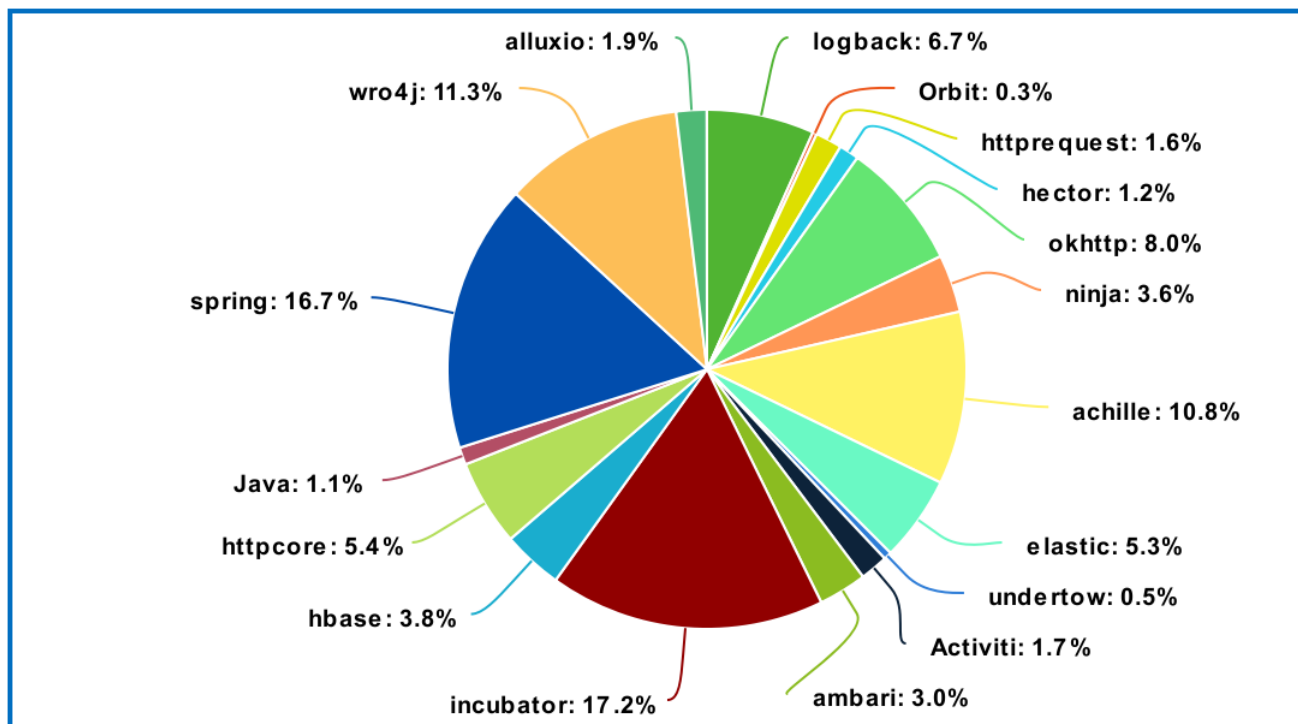
Per la natura del suddetto dataset, è possibile svolgere un'analisi sulle diverse caratteristiche di codifica del singolo caso di test, permettendo quindi un'analisi statistica sulla natura di tali caratteristiche o features, e sul loro impatto a proposito della coverage conseguente.

Costituito da un numero complessivo di quasi 10000 singole unità di test, vengono coinvolti 18 diversi progetti software



```
print(df['nameProject'].unique())  
  
meananalysis x  
/Users/alexlombardi/Desktop/Python/ML/venv/bin/python /Users/alexlombardi/Desktop/Python/ML/venv/bin/python  
['logback' 'orbit' 'http-request' 'hector' 'okhttp' 'ninja' 'Achilles'  
 'elastic-job-lite' 'undertow' 'Activiti' 'ambari' 'incubator-dubbo'  
 'hbase' 'httpcore' 'Java-WebSocket' 'spring-boot' 'wro4j' 'alluxio']
```

**Figura 3.1:** i diversi nomi dei progetti



**Figura 3.2:** distribuzione dei test sui diversi progetti

### 3.1 Descrizione delle features

Production and Test Code Metrics	
Nome	Descrizione
TLOC	Number of lines of code of the test suite
TmcCabe	McCabe cyclomatic complexity, indica la complessità del codice
Lcom2	Lack of Cohesion of Methods version 2, i.e., the percentage of methods that do not access a specific attribute averaged over all attributes in the class.
Lcom5	Lack of Cohesion of Methods version 5, i.e., the density of accesses to attributes by methods.
CBO	Coupling Between Object, i.e., the number of dependencies a class has with other classes[3]
WMC	Weighted Methods per Class, i.e., the sum of the complexities (i.e., McCabe's Cyclomatic Complexity) of all the methods in a class [16]. Note that Chidamber and Kemerer [16] did not define a predefined complexity metric to consider for the computation of WMC. In our case, we opted for the McCabe metric to account for the individual complexity of methods.[3]
RFC	Response For a Class, i.e., the number of methods (including inherited ones) that can potentially be called by other classes[3]
MPC	Message Passing Coupling, measures the numbers of messages passing among objects of the class.
Halstead Vocabulary	The total number of distinct operators and operands in a function
Halstead Length	The total number of operator occurrences and the total number of operand occurrences.
Halstead Volume	Proportional to program size, represents the size, in bits, of space necessary for storing the program.
numCoveredLines	Total number of lines of code covered by the test
executionTime	Running time for the test execution
projectSourceLinesCovered	Total number of production classes covered by each test
hIndexModPerCoverLine	hIndex capturing churn of covered lines in past 5, 10, 25, 50, 75, 100, 500, and 10,000 commits. Each value h indicates that at least h lines were modified at least h times in that period.

Code smells	
Nome	Descrizione
classDataShouldBePrivate	When a class exposes its attributes, violating the information hiding principle.
complexClass	When a class has a high cyclomatic complexity.
functionalDecomposition	When in a class inheritance and polymorphism are poorly used
godClass	When a class has huge dimension and implementing different responsibilities.
spaghettiCode	When a class has no structure and declares long method without parameters.

Text smells	
Nome	Descrizione
Assertion density	percentage of assertion statements in the test code
Assertion roulette	undocumented assertions in the test code
Mystery Guest	Il test presenta materiale esterno[4]
Eager test	Il test analizza più metodi contemporaneamente[4]
Sensitive equality	Il test presenta un confronto sul toString[4]
Resource Optimism	Il test fa uso di risorse esterne potenzialmente non disponibili[4]
Conditional test logic	Il test prevede un if statement condizionale
Fire and forget	Il test lancia attività secondarie in background



## 3.2 Analisi delle features

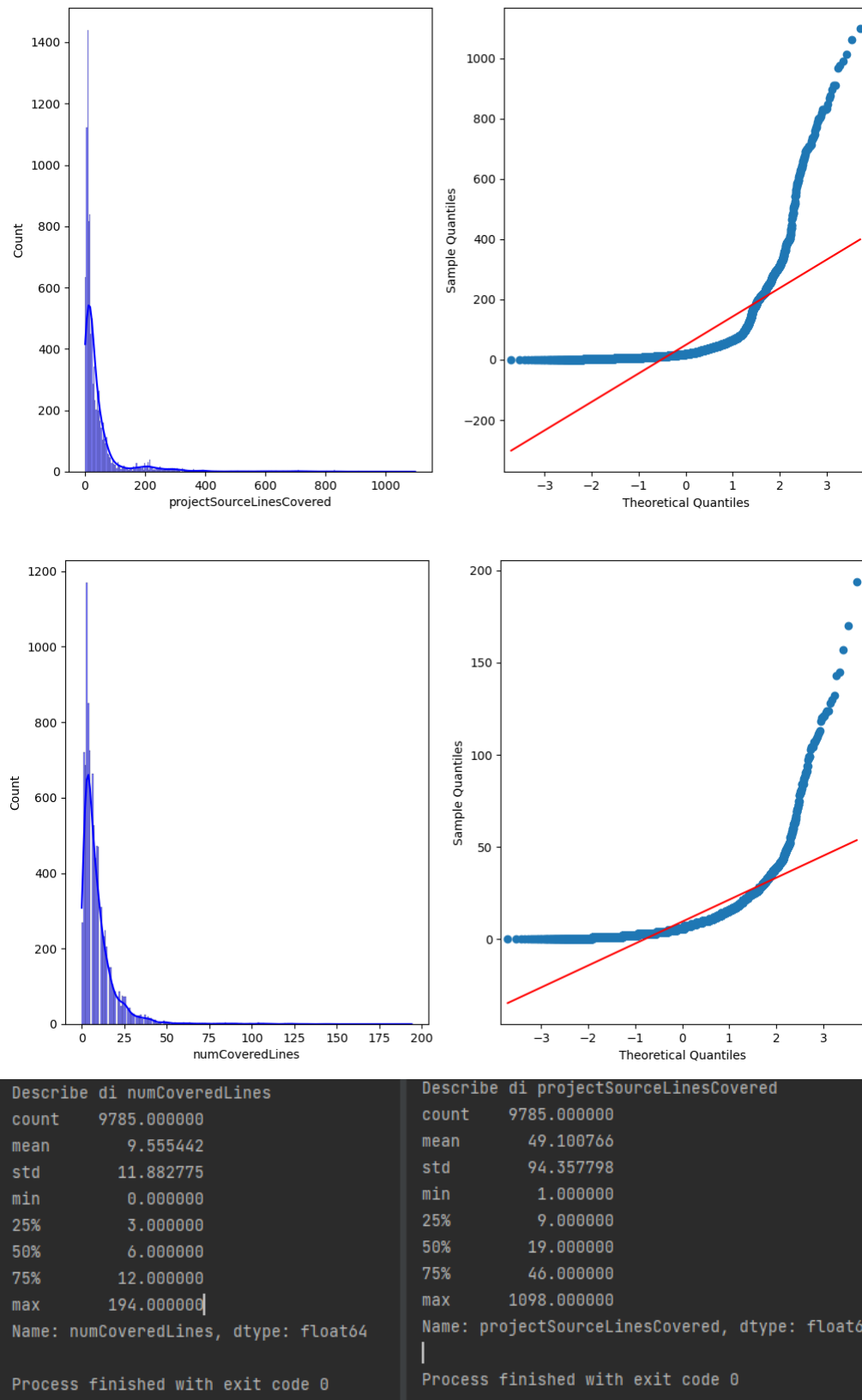
L'obiettivo è quello di andare a costruire un modello tale che, preso in input un test case rappresentato sotto forma di variabili decimali: le con feature statiche, tale modello cercherà di predire la coverage conseguita dal test in questione.

Nello sviluppo di un modello di machine learning, eventualmente si arriverà nella fase di **training**. Nella suddetta fase al modello verranno 'dati in pasto delle informazioni' sicché esso possa comprendere indirettamente gli aspetti e le tematiche costituenti il problema trattato. Nel nostro caso le informazioni sono tutto ciò che è presente all'interno del dataset. Si potrebbe quindi supporre che all'aumentare del quantitativo di informazioni fornite al modello in fase di training, aumenterà di conseguenza anche la comprensione del modello dello scenario e quindi anche le sue performance. Tuttavia questa ipotesi è errata in quanto si, fino ad un certo limite, aumentando la profondità dell'informazione ne risentirà in positivo l'addestramento del modello. Raggiunto questo limite però si incomberà in diverse problematiche spiacevoli, prima fra tutti l'**overfitting** [9]: situazione in cui un modello si adatta troppo bene ai dati di training e, di conseguenza, non può prevedere in modo accurato i dati di test non visualizzati; in altre parole il modello ha un riferimento 'teorico' troppo forte e non riesce ad affrontare una situazione che non ha già visto nel training. Una problematica conseguente ad un training troppo profondo è l'incremento della complessità del modello: per quanto possibile, è nel nostro interesse che il modello sia semplice e comprensibile, pena l'inutilizzabilità del risultato finale.

Per queste motivazioni occorre andare a fare un'analisi delle 38 feature presenti nel dataset per poi selezionarne un sott'insieme: quelle più adatte al training del modello, in altre parole quelle feature che riescono a descrivere il problema in maniera più efficiente rispetto alle altre [8].

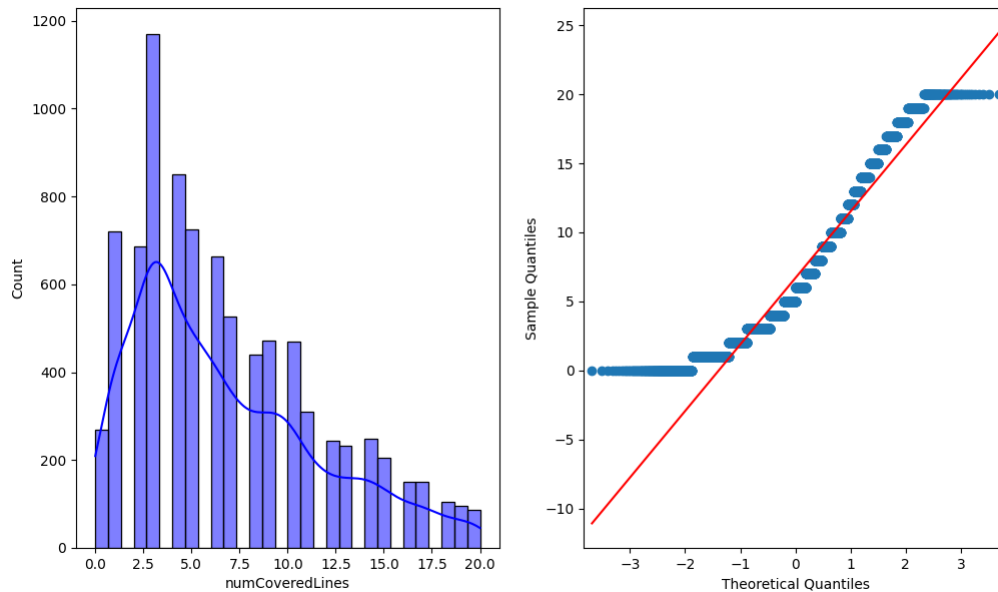
## 3.3 Variabile dipendente

La variabile dipendente ha un ruolo cruciale in quanto esprime in se stessa l'informazione che si sta cercando di far predire al modello. Considerando il paragrafo precedente in cui vengono elencate le diverse features del dataset, si può notare che il ruolo della variabile dipendente ha due possibili candidati: "numCoveredLines" e "projectSourceLinesCovered"



**Figura 3.3:** probability plot e descrizione di "projectSourceLinesCovered" e "numCoveredLines"

Considerando le Figure 3.3 e 3.4 è possibile analizzare la distribuzione dei valori di queste due feature. Per prima cosa occorre notare come in entrambe le feature esistono delle istanze anomale che si discostano dal valore medio, ad esempio per quanto riguarda la variabile "projectSourceLinesCovered" alcune istanze raggiungono valori come 1098 che si discosta molto dal valore medio di 48. Occorre quindi andare a fare una pulizia degli outliers che potrebbero ingannare il modello durante il suo apprendimento.



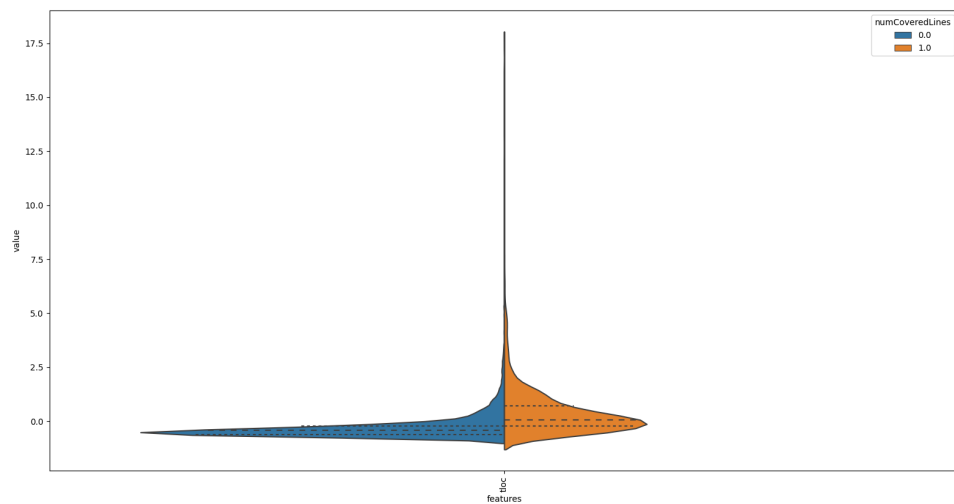
**Figura 3.4:** probability plot di "numCoveredLines" dopo aver rimosso gli outliers

In conclusione durante lo sviluppo del modello sarà presa in considerazione la feature "numcoveredlines" per una semplice questione di attinenza diretta allo scopo di questa tesi: riuscire a predire il numero di righe di codice coperte da un singolo caso di test.

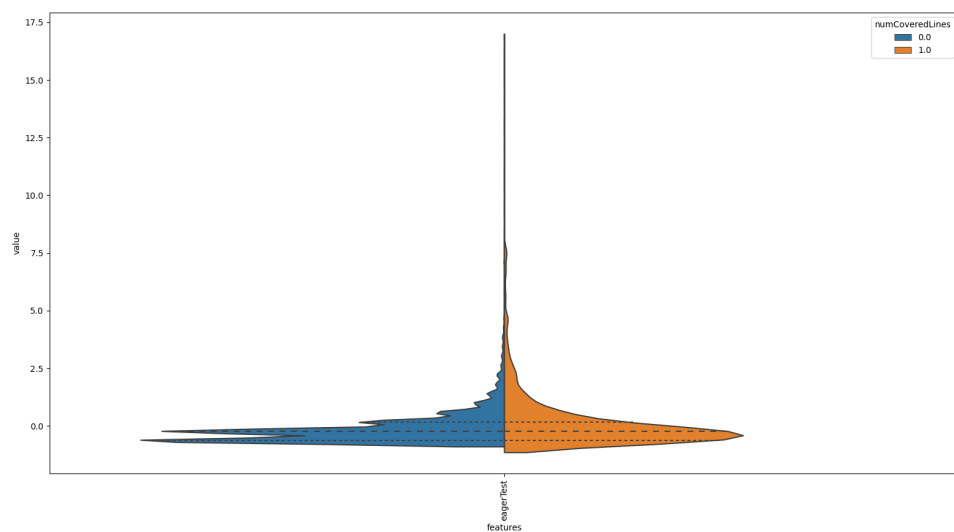
### 3.4 Violin plot

Quello che stiamo cercando è una o più feature sicché i due possibili stati della variabile dipendente, in riferimento alla feature in analisi, assuma valori il più diverso possibili per stato. Quello che stiamo cercando quindi è una feature tale che il suo violin plot corrispondente abbia le righe orizzontali centrali, ovvero quelle rappresentanti la mediana, il più lontane possibili sull'asse delle ordinate. Un buon esempio è quello che troviamo nella Figura 3.5 ovvero la violin plot della feature tloc. Dal grafo è possibile notare che la distribuzione dei valori raggiunti dalle singole istanze presenta uno scarto significativo nei due possibili stati della variabile dipendente, questo è un indizio a proposito della capacità di questa feature a

scindere e a trovare una disgiunzione tra i test soddisfacenti da quelli non soddisfacenti. Quello che invece non stiamo cercando è il caso presente nella Figura 3.6: qui si può notare che le due mediane quasi si incontrano, sintomo che secondo tale feature, i test positivi sono indistinguibili da quelli negativi, si tratta quindi di una feature potenzialmente pessima da dare in training al nostro modello. Altre feature interessanti sono in ordine di potenziale: CBO +++, LOC, RFC, MPC, halsteadVocabulary, halsteadLength, halsteadVolume



**Figura 3.5:** violin plot della feature 'tloc'



**Figura 3.6:** violin plot della feature 'EagerTest'

### 3.5 Indice di correlazione di Pearson

Si sta cercando di dimostrare una correlazione tra la nostra variabile dipendente 'numCoveredLines' e le variabili indipendenti presenti nel dataset. l'indice di correlazione di Pearson tra due variabili statistiche è un indice che esprime un'eventuale relazione di linearità tra esse. Date due variabili statistiche X e Y, l'indice di correlazione di Pearson è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili

The Pearson correlation coefficient [1] measures the linear relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

This function also performs a test of the null hypothesis that the distributions underlying the samples are uncorrelated and normally distributed. (See Kowalski [3] for a discussion of the effects of non-normality of the input on the distribution of the correlation coefficient.) The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. Il test viene computato su Python mediante l'utilizzo della libreria scipy.stats

```
import pandas as pd
from scipy.stats import pearsonr

data = pd.read_csv("csvume/dataset.csv")
list = ['nameProject', 'testCase', "Unnamed: 0", "projectSourceLinesCovered", "isFlaky"]
data = data.drop(list, axis=1)
print(data.columns)
data2 = data.numCoveredLines

for i in range(1, 36):
    data1 = data.iloc[:, i]
    stat, p = pearsonr(data1, data2)
    print("Variabile: " + data.columns[i])
    print('stat={0:.3f}, p={0:.3f}'.format(stat, p))
    if p > 0.05:
        print('Probably independent\n')
    else:
        print('Probably dependent\n')
```

Figura 3.7

<b>Variabile</b>	<b>p value</b>	<b>esito probabilistico</b>
tloc	0.543	dependent
tmcCabe	0.190	dependent
assertionDensity	0.189	dependent
assertionRoulette	0.359	dependent
mysteryGuest	0.113	dependent
eagerTest	0.091	dependent
sensitiveEquality	0.058	dependent
resourceOptimism	-0.004	Independent
conditionalTestLogic	0.196	dependent
fireAndForget	0.2	dependent
loc	0.051	dependent
lcom2	0.005	Independent
lcom5	0.002	Independent
cbo	0.092	dependent
wmc	0.063	dependent
rfc	0.082	dependent
mpc	0.088	dependent
halsteadVocabulary	0.088	dependent
halsteadLength	0.058	dependent

Variabile	p value	esito probabilistico
halsteadVolume	0.057	dependent
classDataShouldBePriv	-0.025	dependent
complexClass	0.022	dependent
functionalDecomp	-0.039	dependent
godClass	0.024	dependent
spaghettiCode	0.018	Independent
ExecutionTime	0.089	dependent
hIndexMod..dLine5	0.277	dependent
hIndexMod..dLine10	0.233	dependent
hIndexMod..dLine25	0.200	dependent
hIndexMod..dLine50	0.158	dependent
hIndexMod..dLine75	0.128	dependent
hIndexMod..dLine100	0.145	dependent
hIndexMod..dLine500	0.351	dependent
hIndexMod..dLine10000	0.296	dependent

In questo capitolo si scenderà nel dettaglio in merito alla costruzione del modello di regressione. Tale modello prenderà in input un insieme di feature indipendenti appartenenti ad un generico test case e restituirà in output una predizione sul numero di righe di codice coperte dal suddetto test case.'

```
df = pd.read_csv("dataset.csv")
df = df[df['nameProject'].str.match('logback')]
list = ['nameProject', 'testCase', "Unnamed: 0", "projectSourceLinesCovered"]
y = df.numCoveredLines
df = df.drop(list, axis=1)
# print(df.columns)
# split data train 70 % and test 30 %
x_train, x_test, y_train, y_test = train_test_split(df, y, test_size=0.3, random_state=42)
y_train = np.log1p(y_train)

matplotlib.rcParams['figure.figsize'] = (12.0, 6.0)
prices = pd.DataFrame({"price": x_train["numCoveredLines"], "log(price + 1)": np.log1p(x_train["numCoveredLines"])}).
prices.hist()
plt.show()
```

**Figura 4.1**

In un primo momento vengono rimosse dal dataset le feature non interessanti ai fini della predizione (ad esempio l'id o il nome del progetto) così come le feature ritenute 'troppo interessanti' quali projectSourceLinesCovered; con il metodo train-test-split della libreria sklearn si partiziona agevolmente il dataset in due sott'insiemi uno per il training ed uno per il testing.



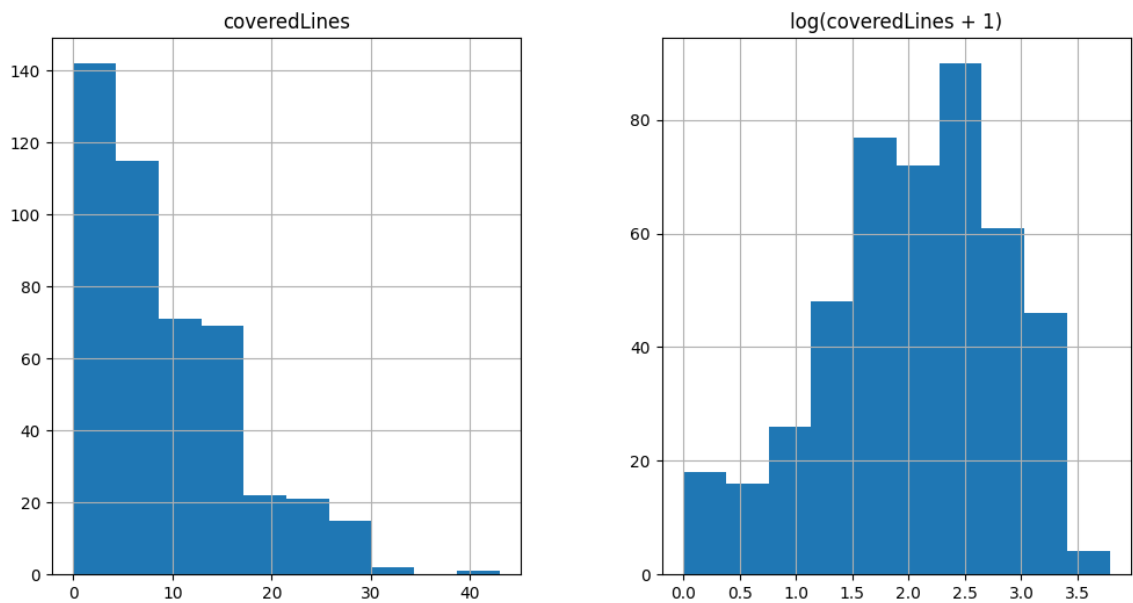


Figura 4.2

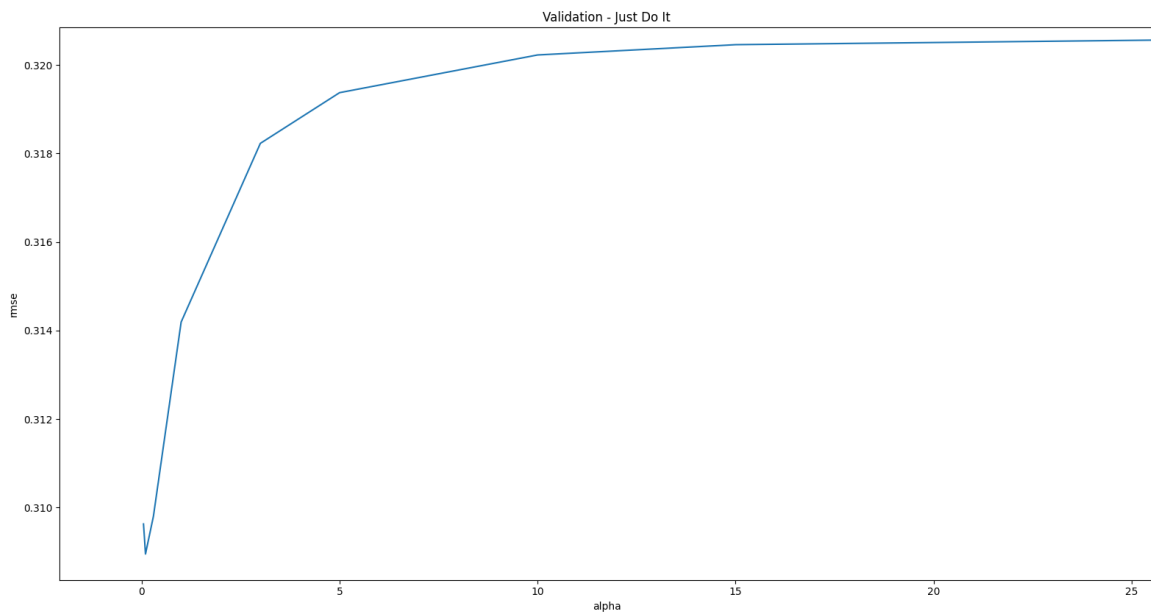
Andando ad analizzare la distribuzione della nostra variabile dipendente è evidente come la maggioranza dei test case presenti nel dataset abbiano una tendenza ad assumere un valore nel range 0-10, il che implica una distribuzione non lineare dei valori raggiunti dalla variabile dipendente nel dataset. Al fine di agevolare l'apprendimento del nostro modello, c'è interesse nel rendere il comportamento della variabile dipendente più lineare, è possibile fare questo applicando banalmente il logaritmo all'intera colonna `coveredLines` (la nostra variabile dipendente) la quale assumerà valori distribuiti in maniera più equilibrata verso il valore medio.

In seguito verrà descritta l'implementazione di tre diversi modelli di regressione. A dire la verità i tre modelli saranno concettualmente identici, l'elemento di discordanza è la tecnica di regolarizzazione adottata ovvero 'lasso', 'ridge' e XGB. I tre modelli saranno confrontati inizialmente tramite la misura RMSE alias 'Radice dell'errore quadratico medio' anche se più semplicemente ed intuitivamente basterebbe andare a fare il confronto tra il valore assoluto della differenza tra il valore predetto ed il valore reale.

## 4.1 Ridge model

Il primo modello adotterà il metodo di regolarizzazione noto come Ridge. Tale metodo di regolarizzazione ha la peculiarità di ricorrere ad un parametro di configurazione definito  $\alpha$  che inciderà sulla flessibilità del modello. Maggiore sarà la regolarizzazione minore sarà

la probabilità di andare in overfit, ciononostante il modello perderà di flessibilità incombendo nel rischio di non riuscire a captare tutte le informazioni presenti nei dati.



**Figura 4.3**

Dal grafico si evince che il coefficiente ideale da assegnare ad  $\alpha$  è poco superiore allo 0, superato tale valore la radice dell'errore quadratico medio non farà altro che aumentare. Trovato il coefficiente  $\alpha$ , una volta effettuato il training sul modello ed eseguite le predizioni sulle istanze del dataset dedicate al testing, si avrà un RMSE del **0.6272**. In altre parole avremo uno scarto assoluto medio (MAE) di 6.15 linee di codice di differenza tra il valore predetto e quello effettivo; analizzando le predizioni però si evince che lo scarto tra valore predetto ed effettivo è direttamente proporzionale alla grandezza del valore reale. Andando a scartare tutti i test le cui linee di codice corrispondenti superino 20 loc, lo scarto tra valore predetto e reale sarà di 3.22.

## 4.2 Lasso model

Il secondo modello adotterà la tecnica di regressione nota come Lasso. Questa tipologia di regressione ha la peculiarità di andare automaticamente a determinare quali sono le feature interessanti o meno ai fini dell'addestramento del modello. Sebbene sia una tecnica che potrebbe offrire spunti interessanti ai fini di questo articolo, non si andrà troppo nel dettaglio in questo paragrafo sicché il valore dell'RMSE è pari a **0.6279** leggermente superiore a quello

del modello precedente e anche qui lo scarto assoluto medio tra il valore predetto ed effettivo è di 6.15. In conclusione si evince che tra Ridge e Lasso come tecniche di regolarizzazione il risultato è funzionalmente identico.

### 4.3 Extreme Gradient Boosting (XGBoost)

Il terzo modello è stato costruito ricorrendo alla libreria open source XGBoost adatta allo sviluppo di modelli di regressione supervisionati. Tale modello adotta la tecnica dell'**ensemble**: nella fase di apprendimento saranno combinati più modelli individuali (base learners) ognuno di essi farà la sua predizione. Esistono diverse tipologie di apprendimento ensemble, in questo caso verrà adottata la tecnica del **boosting** ovvero ciasun base learner influirà sulla predizione finale con un certo peso. Il coefficiente del peso viene calcolato in base all'errore commesso in fase di learning.

Una caratteristica è il formato del dataset utilizzato dal modello ovvero il DMatrix. Una tipologia di struttura dati che va ad incrementare le performance e l'efficienza.

Un modello XG durante la fase di training analizza la complessità di ogni base learner, l'obiettivo è quello di ottenere come risultato un modello semplice ed accurato, qualora tramite una loss function uno dei diversi base learner divenisse troppo complicato, per prevenire il rischio di overfitting vengono combinate tecniche di regolarizzazione quali LASSO e Ridge.

Il terzo modello offre un leggero miglioramento se confrontato con i suoi predecessori, riesce a predire la variabile dipendente con un RMSE del 0.4825 e quindi uno scarto medio assoluto di 3.45 linee di codice oppure 2.33 considerando solo i casi dal LOC inferiore a 25

### 4.4 Random Forest Regression

<https://towardsdatascience.com/random-forest-regression-5f605132d19d> L'ultimo modello è il più preciso in quanto è in grado di predire il numero delle covered lines con uno scarto assoluto medio di 2.31 linee di codice. E' stato usato il modulo sklearn per lo sviluppo e il training del random forest regression model, nello specifico la funzione RandomForestRegressor.

```
#RIDGE
model_ridge = Ridge(alpha = 0.05).fit(x_train, y_train)
ridge_preds = np.exp1(model_ridge.predict(x_test))
predizioni1 = pd.DataFrame({"id":x_test.id, "coveredLines":ridge_preds})
predizioni1.to_csv("ridge.csv", index = False)

#LASSO
model_lasso = LassoCV(alphas = [1, 0.1, 0.001, 0.0005]).fit(x_train, y_train)
print(rmse_cv(model_lasso).mean())
coef = pd.Series(model_lasso.coef_, index = x_train.columns)
lasso_preds = np.exp1(model_lasso.predict(x_test))
predizioni1 = pd.DataFrame({"id":x_test.id, "coveredLines":lasso_preds})
predizioni1.to_csv("lasso.csv", index = False)

#XGB
import xgboost as xgb
dtrain = xgb.DMatrix(x_train, label = y_train)
params = {"max_depth":2, "eta":0.1}
model = xgb.cv(params, dtrain, num_boost_round=500, early_stopping_rounds=100)

model_xgb = xgb.XGBRegressor(n_estimators=360, max_depth=2, learning_rate=0.1)
model_xgb.fit(x_train, y_train)
xgb_preds = np.exp1(model_xgb.predict(x_test))
predizioni1 = pd.DataFrame({"id":x_test.id, "coveredLines":xgb_preds})
predizioni1.to_csv("xgb.csv", index = False)
```

Figura 4.4

---

### Considerazioni e Risultati

---

Notare come risultato un MAE del 2.31 è frutto del training eseguito su tutto il dataset per intero, andando a partizionare il dataset e a considerare i singoli progetti che lo compongono si otterranno i risultati descritti nella tabella che segue.

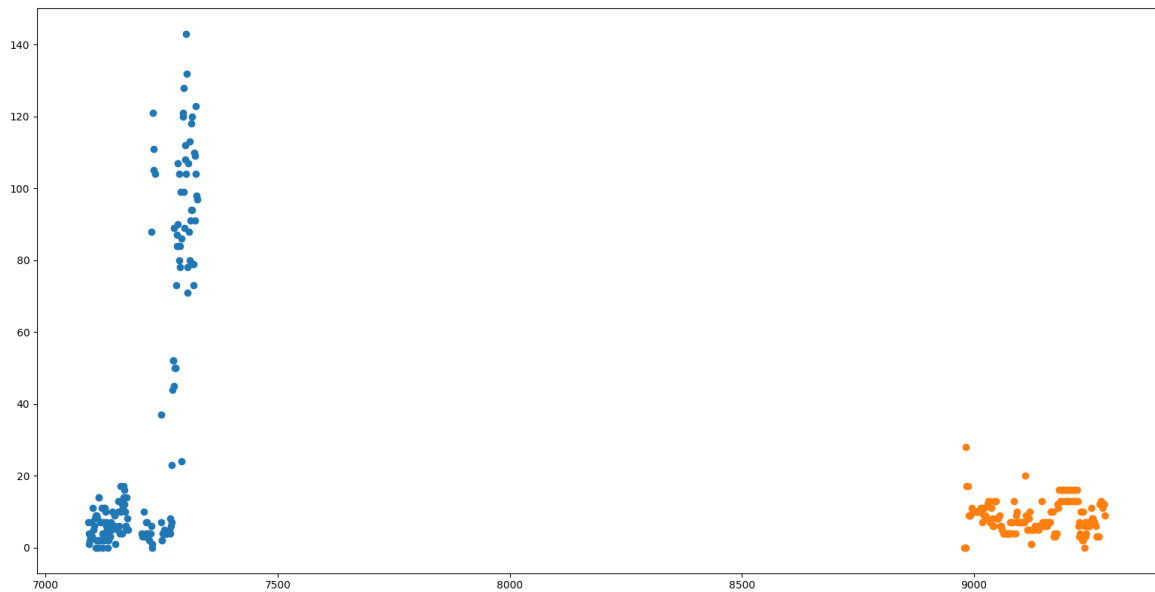
Progetto	Scarto assoluto medio	Percentuale del DS
logback	2.882	6.7%
orbit	4.152	0.3%
http-request	1.167	1.6%
hector	4.303	1.2%
okhttp	1.593	8.0%
ninja	1.117	3.6%
Achilles	1.220	10.8%
elastic-job-lite	1.376	5.3%
undertow	10.093	0.5%
Activiti	6.065	1.7%
ambari	6.361	3%
incubator-dubbo	1.968	17.2%
hbase	4.774	3.8%
httpcore	2.487	5.4%
Java-WebSocket	2.406	1.1%
spring-boot	1.606	16.7%
wro4j	1.645	11.3%
alluxio	3.304	1.9%

Dalla tabella si evince che lo scarto assoluto medio è inversamente proporzionale alla percentuale del progetto nel dataset. In altre parole, maggiore è il numero dei singoli test case associati al singolo progetto, maggiore sarà la capacità predittiva del modello nel determinare la coverage associata ad un nuovo test case.

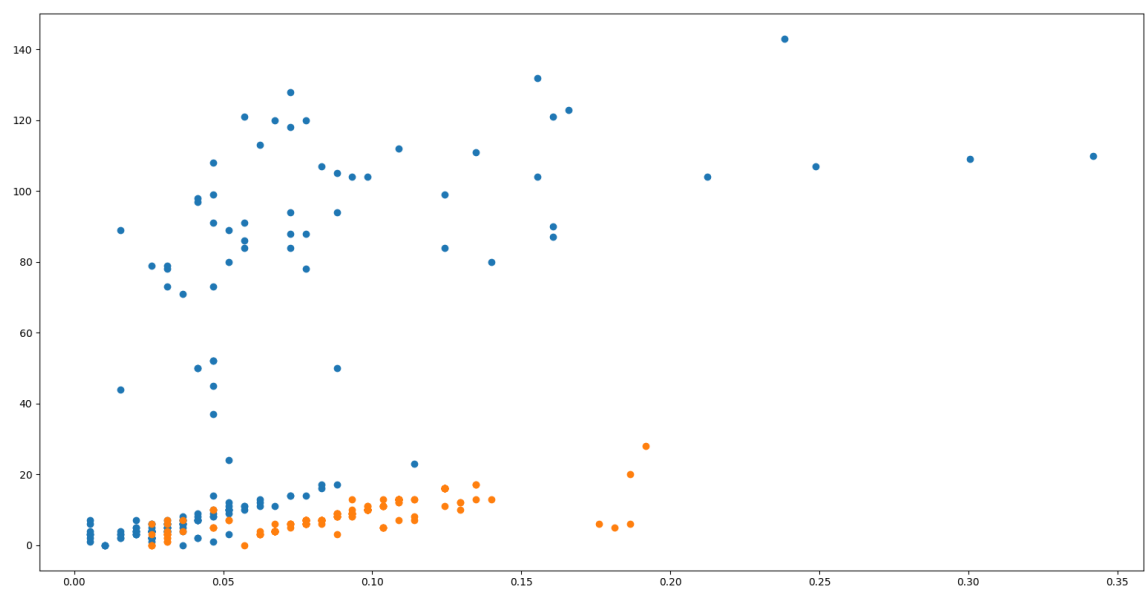
Ciononostante ci sono comunque degli outliers a questa supposizione, considerando i due progetti 'http-request' e 'Activiti' la leggera differenza nella percentuale non giustifica la forte differenza nello scarto assoluto medio dei rispettivi, ne consegue che per alcuni progetti è più facile predire la codecoverage rispetto che ad altri, ma perché?

nella figura 5.1 viene descritta la distribuzione dei coefficienti raggiunti della variabile dipendente numCoveredLines nei due progetti 'Activiti' in blu e 'http-request' in arancio. Dall'immagine è possibile notare che nel primo progetto i valori raggiunti dalla variabile dipendente hanno una varianza superiore rispetto che a quelli nel secondo progetto.

nella figura 5.2 è mostrato un boxplot tra la variabile indipendente più significativa trovata durante la costruzione del modello 'tloc' e la variabile dipendente 'numCoveredLines'. E'



**Figura 5.1:** scatter plot distribuzione variabile dipendente arancio = progetto http blu = progetto  
attività



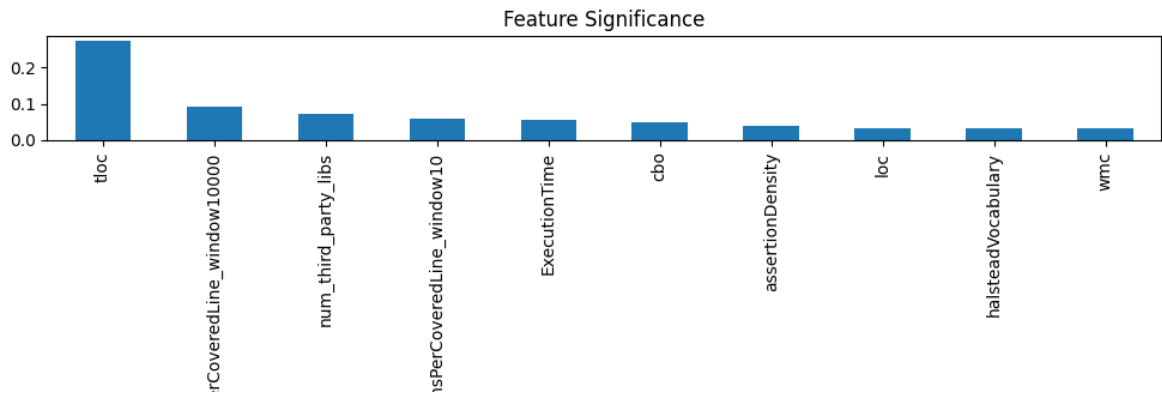
**Figura 5.2**

evidente come vi è una linearità più definita tra le variabili del progetto in arancione 'http-request' rispetto che alla sua controparte. Si suppone che questo comportamento del progetto 'Attività' è dovuto a istanze di test unit con una particolarità descritta da degli outliers nel data, nello specifico un picco nei valori raggiunti dalla variabile dipendente non descritto nelle variabili indipendenti considerate.

Andando a rimuovere gli outliers su tutto il dataset l'accuratezza del modello più preciso

raggiunge uno scarto assoluto medio di **1.5**

Di seguito è proposta un'analisi nell'impatto che le diverse variabili indipendenti hanno avuto nell'addestramento del modello ricorrendo al metodo `feature_importances` della libreria `sklearn`



**Figura 5.3**

si evince che la variabile indipendente più significativa nel training del modello è TLOC ovvero il quantitativo di linee di codice da cui è costituito il singolo test case. Andando a rieffettuare il training del modello con il medesimo dataset ma senza la colonna 'TLOC' si ottiene uno scarto assoluto medio di **1.72**, un incremento dello 0,2 rispetto al valore ottenuto in precedenza. Ciò implica che sì la variabile TLOC è significativa nella descrizione dei singoli test case, ma non fondamentale o almeno non è l'unica in grado di fornire una descrizione della realtà trattata al modello.



---

Ringraziamenti

---

INSERIRE RINGRAZIAMENTI QUI

---

## Bibliografia

---

- [1] Dhaya Sindhu Battina. Artificial intelligence in software test automation: A systematic literature review. *International Journal of Emerging Technologies and Innovative Research* ([www.jetir.org](http://www.jetir.org) | UGC and issn Approved), ISSN, pages 2349–5162, 2019. (Citato a pagina 1)
- [2] Boris Beizer. *Software Testing Techniques (2nd Ed.)*. Van Nostrand Reinhold Co., USA, 1990. (Citato a pagina 1)
- [3] S.R. Chidamber and C.F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994. (Citato a pagina 11)
- [4] Arie Van Deursen, Leon Moonen, Alex Bergh, and Gerard Kok. Refactoring test code. In *Proceedings of the 2nd International Conference on Extreme Programming and Flexible Processes in Software Engineering (XP2001)*, pages 92–95, 2001. (Citato a pagina 12)
- [5] David P Doane and Lori E Seward. Measuring skewness: a forgotten statistic? *Journal of statistics education*, 19(2), 2011.
- [6] B. Hailpern and P. Santhanam. Software debugging, testing, and verification. *IBM Systems Journal*, 41(1):4–12, 2002. (Citato a pagina 1)
- [7] Cem Kaner. Exploratory testing. In *Quality assurance institute worldwide annual software testing conference*, pages 1–14, 2006.
- [8] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992. (Citato a pagina 13)

- [9] Erick Odhiambo Omuya, George Onyango Okeyo, and Michael Waema Kimwele. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 174:114765, 2021. (Citato a pagina 13)
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [11] Strategic Planning. The economic impacts of inadequate infrastructure for software testing. *National Institute of Standards and Technology*, page 1, 2002. (Citato a pagina 1)
- [12] Valeria Pontillo, Fabio Palomba, and Filomena Ferrucci. Toward static test flakiness prediction: A feasibility study. In *Proceedings of the 5th International Workshop on Machine Learning Techniques for Software Quality Evolution, MaLTESQuE 2021*, page 19–24, New York, NY, USA, 2021. Association for Computing Machinery. (Citato a pagina 9)
- [13] Wikipedia contributors. Pearson correlation coefficient — Wikipedia, the free encyclopedia, 2022. [Online; accessed 18-July-2022].

### Siti Web consultati

- Wikipedia – [www.wikipedia.org](http://www.wikipedia.org)