

# Relazione Progetto Statistica Numerica

## 1. Selezione dataset

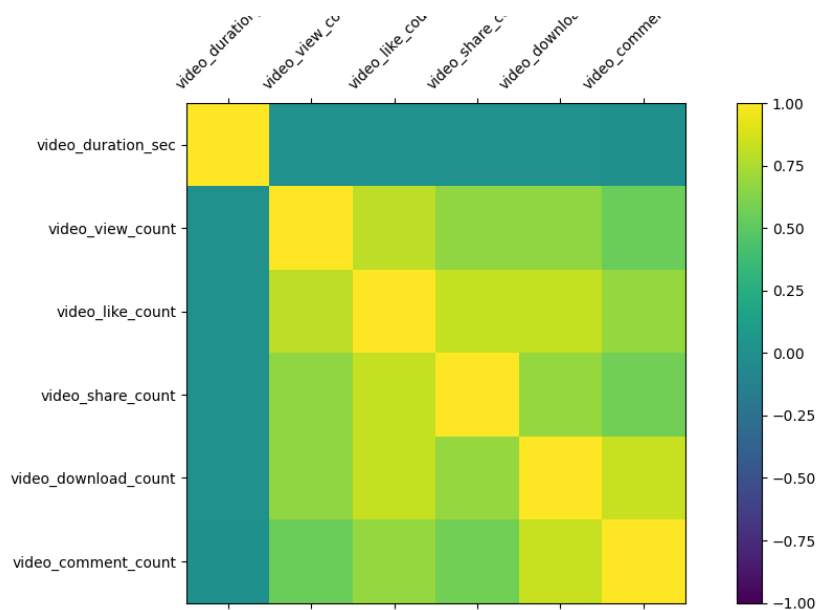
È stato scelto il seguente dataset: <https://www.kaggle.com/datasets/yakhyojon/tiktok>. Il dataset contiene circa 19mila entry riguardo a video pubblicati su TikTok, divisi in "affermazioni" e "opinioni". Con opinione si intende a un pensiero di un singolo o di un gruppo, mentre un'affermazione è un'informazione senza fonte o con fonte non verificata. Per ogni video si ha inoltre il numero di visualizzazioni, like, commenti, condivisioni e durata. Informazioni presenti sul dataset ma non utilizzate per questo progetto sono lo status dell'account (verificato o non), ID del video e le prime frasi del video.

## 2. Pre-processing

Sono state eliminate le colonne relative al numero della entry e la trascrizione del video. Anche tutte le righe contenenti valori invalidi (nel dataset indicati da "" o "#N/A"). La funzione `isna().any()` conferma che non ci sono valori invalidi nelle colonne del dataset.

Variabili non numeriche sono state convertite in variabili categoriche. Sono anche fatti controlli per assicurarsi che non siano presenti valori negativi all'interno del dataset. Dopo aver ripulito il dataset ci si ritrova con 9608 affermazioni e 9476 opinioni.

## 3. Exploratory Data Analysis



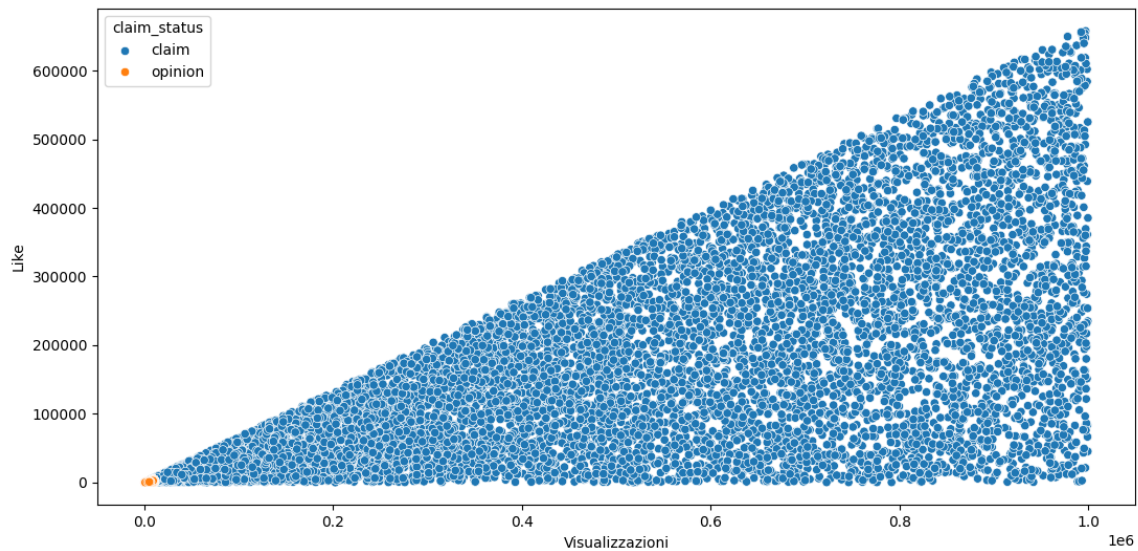
Matrice di correlazione tra le variabili numeriche del dataset.

Si nota come la durata del video non influenza visualizzazioni, like e commenti, dato che ha un valore attorno allo 0.

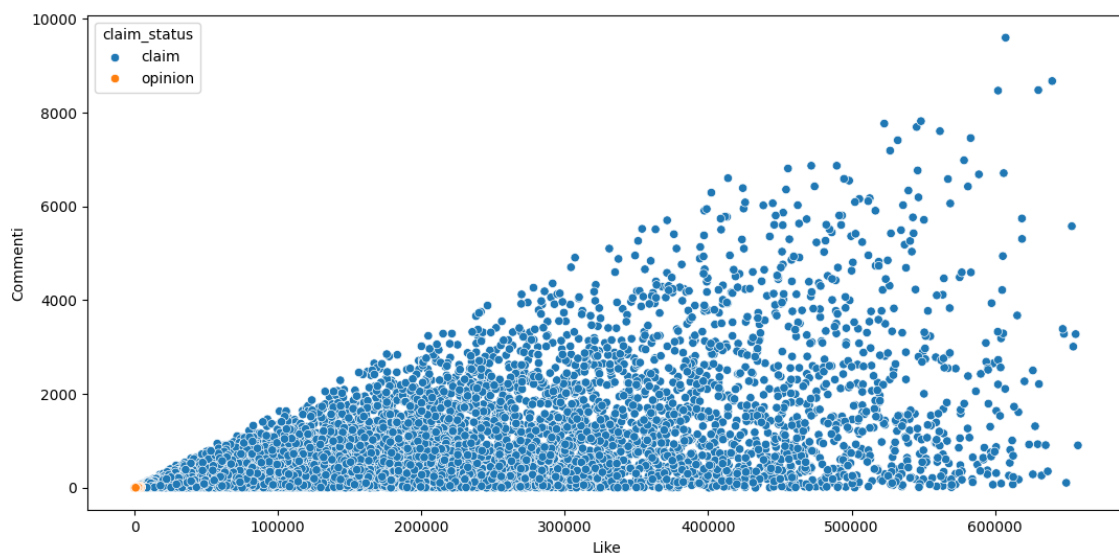
Le visualizzazioni hanno una correlazione positiva con le altre metriche: questo indica che il numero di visualizzazioni influenza fortemente le interazioni con il video (like, commenti,...). Lo stesso vale per il numero di commenti, condivisioni, download e commenti.

Non ci sono metriche il cui aumento causa una riduzione delle altre. Si nota anche l'assenza di valori correlati negativamente, ovvero valori che diminuiscono all'aumentare di altri.

In conclusione possiamo affermare che le interazioni di un video sono direttamente proporzionali al numero di visualizzazioni e viceversa.



In questo scatterplot emerge la forte relazione di proporzionalità diretta tra visualizzazioni e like. Inoltre si nota come i video categorizzati come opinione abbiano molte poche visualizzazioni rispetto ad affermazioni, nonostante il numero di "opinion" e "claim" nel dataset sia quasi lo stesso.



In questo grafico si nota che la densità dei punti diminuisce spostandosi verso destra e verso l'alto, indicando che i video con un altissimo numero di like e commenti sono meno comuni (o con numero di commenti paragonabile a quello delle visualizzazioni). Tuttavia anche in questo caso è presente una correlazione positiva.

## 4. Splitting

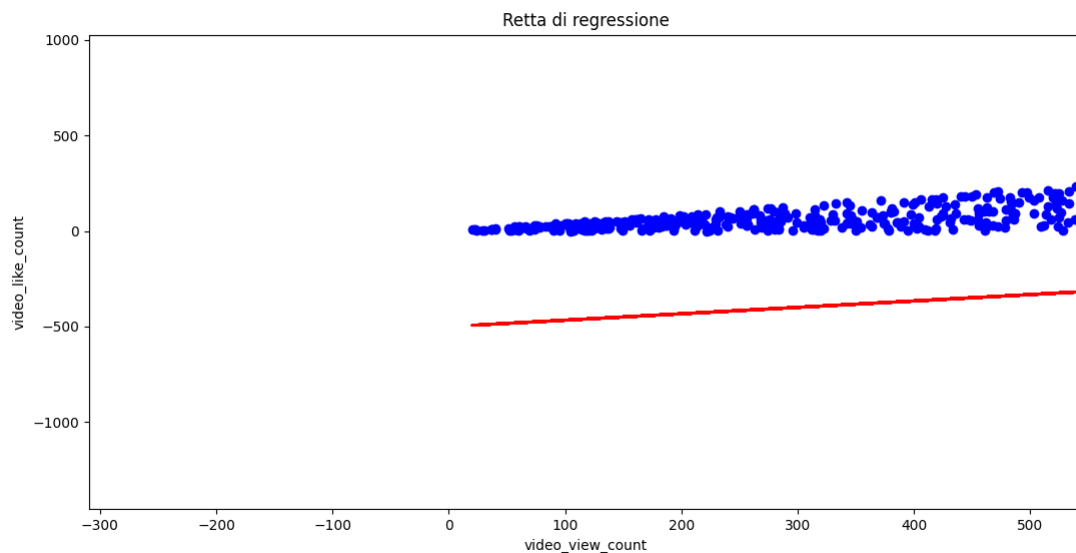
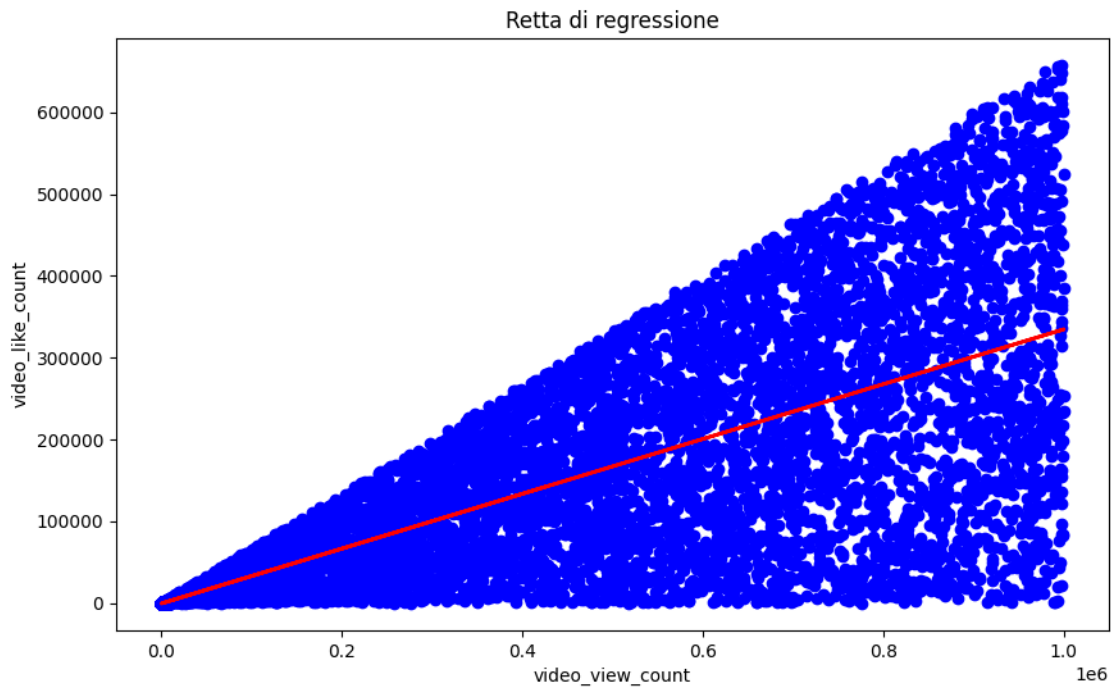
Il dataset è stato diviso in training set (70%), test set (15%) e validation set (15%)

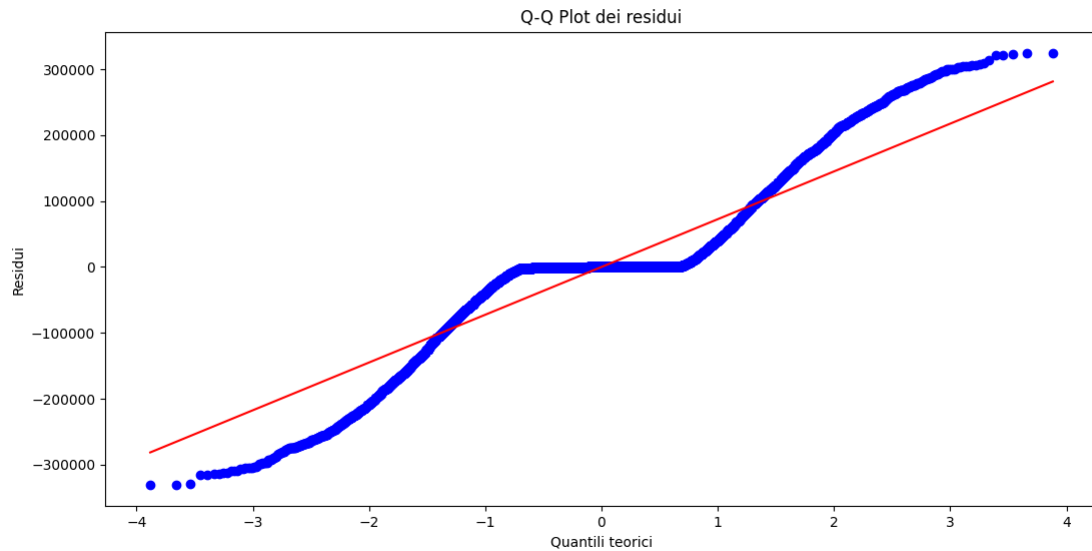
## 5. Regressione

Regressione con  $X = \text{video\_view\_count}$  e  $y = \text{video\_like\_count}$  (fortemente correlati)

- Coefficienti stimati:  $[0.33538031]$ . Questo valore indica la pendenza della retta di regressione. Più questo valore tende a 1, maggiore sono correlate le due variabili.
- Intercetta:  $-500.6339385318279$ . Il dataset fornito non contiene entry con 0 visualizzazioni. Dato che l'intercetta richiede che  $X = 0$ , non può fare predizioni corrette per quel valore in quanto non ha dati sufficienti.

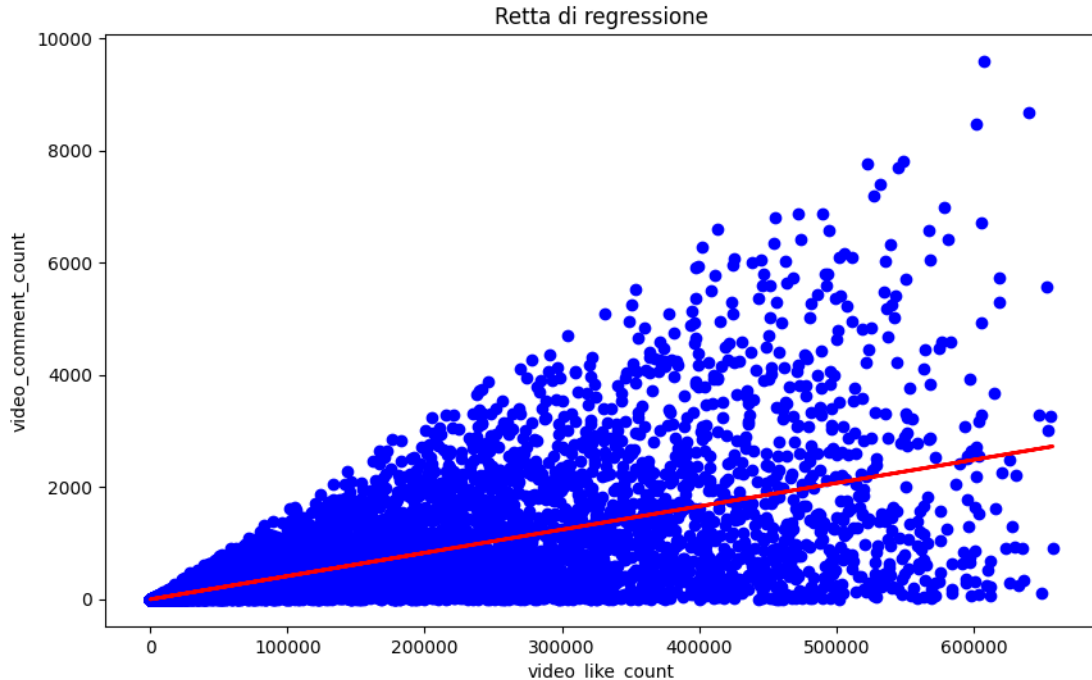
- Coefficiente  $R^2$ : 0.6481893265330756. Questo valore indica che circa il 64.8% della variabilità nei like dipende dalle visualizzazioni. Ci si poteva già aspettare un esito del genere sapendo che c'è forte correlazione tra visualizzazioni e like.
- Errore quadratico medio (MSE): 6394487239.400973. Indica che c'è una grande discrepanza tra i valori predetti dal modello e i valori effettivi, come notato osservando il valore errato dell'intercetta.
- Test di Shapiro-Wilk su un sotto campione:
  - statistica: 0.8148850766554967. I dati sono per l'81% simili a una distribuzione normale.
  - p-value:  $1.0402750750950724 \times 10^{-55}$ . C'è una probabilità molto bassa che i dati siano distribuiti normalmente, quindi si rifiuta l'ipotesi di normalità  $H_0$ .





Regressione con  $X=\text{video\_like\_count}$  e  $y=\text{video\_comment\_count}$  (debolmente correlati)

- Coefficienti stimati: [0.00414611]. Minuscola pendenza della retta, correlazione tra le due variabili molto bassa.
- Intercetta: 2.98502114940203. Questo valore è plausibile perché ci sono video con 0 like sui quali il modello è stato addestrato.
- Coefficiente  $R^2$ : 0.4719091969412126
- Errore quadratico medio (MSE): 349645.55347023753
- Test di Shapiro-Wilk su un sotto campione
  - statistica: 0.6357944320720752
  - p-value:  $1.9487556479935565 \times 10^{-68}$ , si rifiuta l'ipotesi di normalità  $H_0$ .



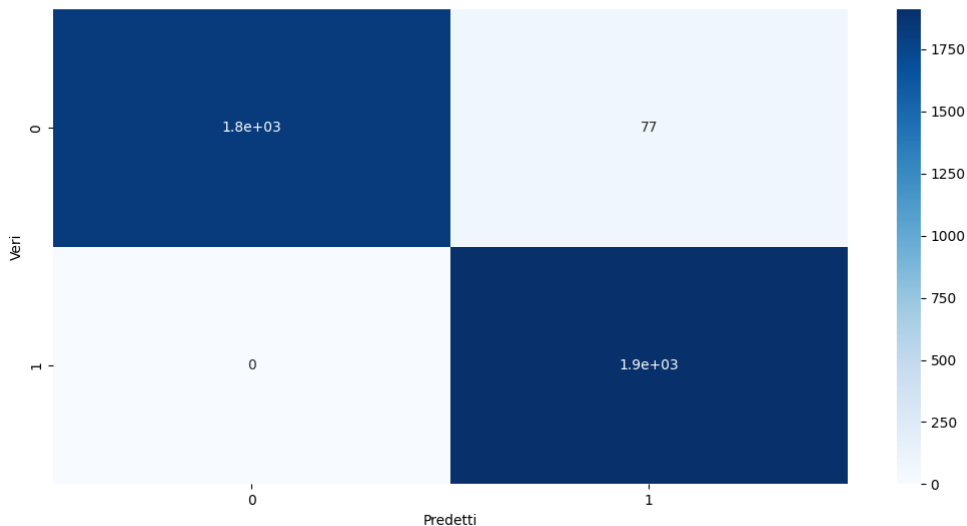
## 6. Addestramento del modello

Si addestra il modello con regressione logistica per predire, dato il numero di visualizzazioni, like, commenti e condivisioni se un video è un'affermazione o un'opinione. Le  $X$  sono state standardizzate per far convergere l'algoritmo e migliorare la performance del modello.

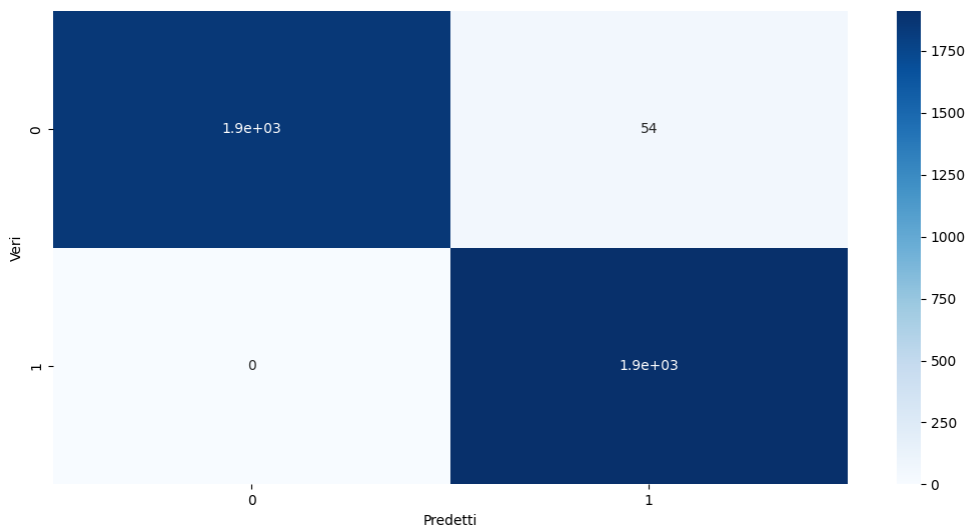
Il report di classificazione afferma che c'è una precisione del 100% nel predire affermazioni, e del 96% per quanto

riguarda le opinioni. Il 98% Delle predizioni totali sono risultate corrette.  
Guardando la matrice di confusione si nota che:

- 1828 è il numero di veri positivi per la classe affermazione.
- 77 è il numero di falsi negativi per la classe affermazione.
- 0 è il numero di falsi positivi per la classe opinione.
- 1912 è il numero di veri positivi per la classe opinione.



Con la Support Vector Classification (SVC) lineare, i risultati sono leggermente migliori. La precisione delle affermazioni resta al 100%, mentre quella delle opinioni è del 97%, portando così a una precisione complessiva del 99%.



## 7. Hyperparameter Tuning

Per trovare il grado con accuratezza migliore in una SVC poly, si eseguono più operazioni fit aumentando il grado e tenendo traccia delle varie precisioni. Si nota che il grado che produce il migliore è il primo, con risultati pressoché identici alla regressione logistica (precisione del 98,008%).

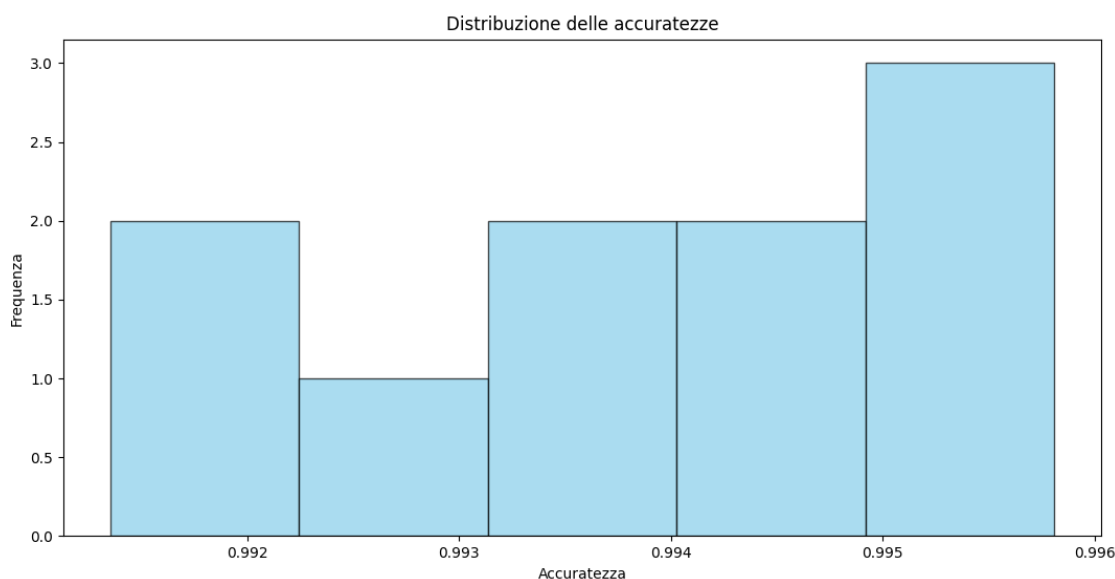
## 8. Valutazione della performance

In conclusione possiamo affermare che il modello che produce i valori più vicini al test set è l'SVC lineare, seguito da SVC poly di grado 1 e regressione logistica. Gradi maggiori di 1 con SVC poly hanno le seguenti accurtezze: 0.9041131778883941, 0.9779931883678281, 0.8467382761330888, 0.9753733298401887.

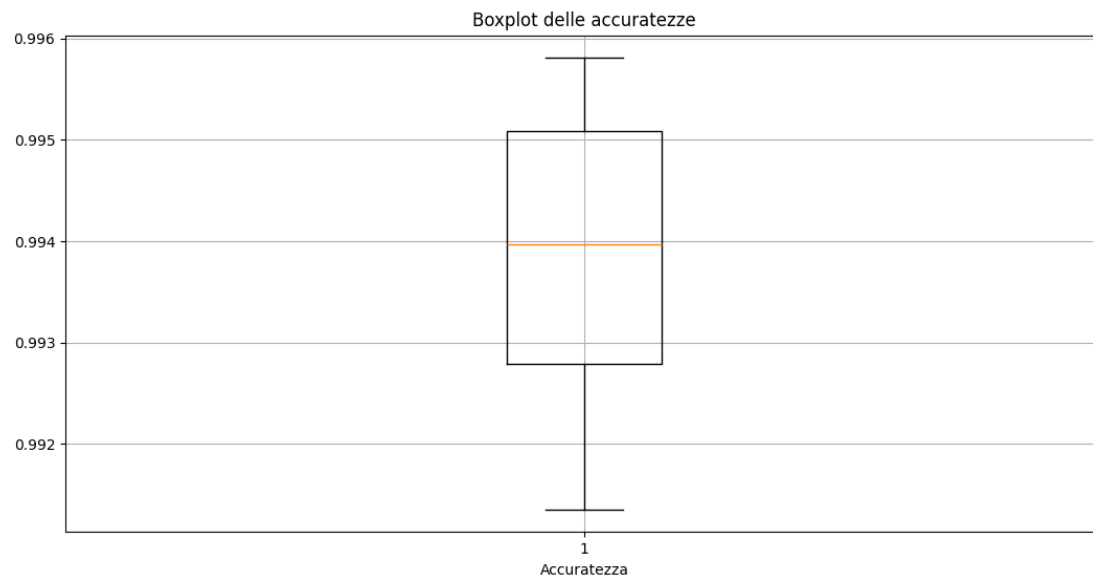
## 9. Studio statistico sui risultati della valutazione

Eseguendo la fase di addestramento 10 volte con la regressione logistica usando test e train set casuali ogni volta si hanno i seguenti risultati.

Iterazione	Errore di clssificazione (ME)	Tasso di errore di classificazione (MR)	Accuratezza (ACC)
1	21	0.005501702908042965	0.994498297091957
2	16	0.004191773644223212	0.9958082263557768
3	18	0.004715745349751114	0.9952842546502488
4	18	0.004715745349751114	0.9952842546502488
5	28	0.007335603877390621	0.9926643961226094
6	22	0.005763688760806916	0.9942363112391931
7	32	0.008383547288446425	0.9916164527115536
8	26	0.006811632171862719	0.9931883678281372
9	24	0.006287660466334818	0.9937123395336652
10	33	0.008645533141210375	0.9913544668587896



L'accuratezza media è 0.9937647367042179, che è molto simile alla mediana: 0.9939743253864292. Questo suggerisce che le accurtezze non sono influenzate da valori estremi (perché la media è molto sensibile agli outlier e la mediana no). Una deviazione standard dello 0.15% indica che c'è una bassa variazione nei valori delle precisioni, dovuta a un modello corretto e valori estremi assenti. Dato che l'istogramma non è simmetrico, si può affermare che la distribuzione dei residui non è normale.



Questi valori sono evidenziati dal boxplot, che con la linea arancione indica la mediana. Il rettangolo nel boxplot rappresenta l'intervallo interquartile, che è l'intervallo tra il primo e il terzo quartile. I baffi indicano l'intervallo complessivo dei dati, escludendo eventuali outlier.

Intervallo di confidenza con  $\alpha = 0.05$  :  $[0.9928554225357857, 0.9946740508726502]$ .

L'intervallo di confidenza indica l'intervallo dove ci si aspetta di trovare il vero valore medio dell'accuratezza con un livello di confidenza del 95%, tenendo conto di variazioni dei dati nelle diverse iterazioni.

## 10. Conclusioni

Si può affermare che data la tendenza delle opinioni ad avere visualizzazioni e altri indici di interazione molto bassi, fare predizioni corrette per questa classe è molto facile per qualsiasi modello.