

Relazione Progetto Statistica Numerica

1. Selezione dataset

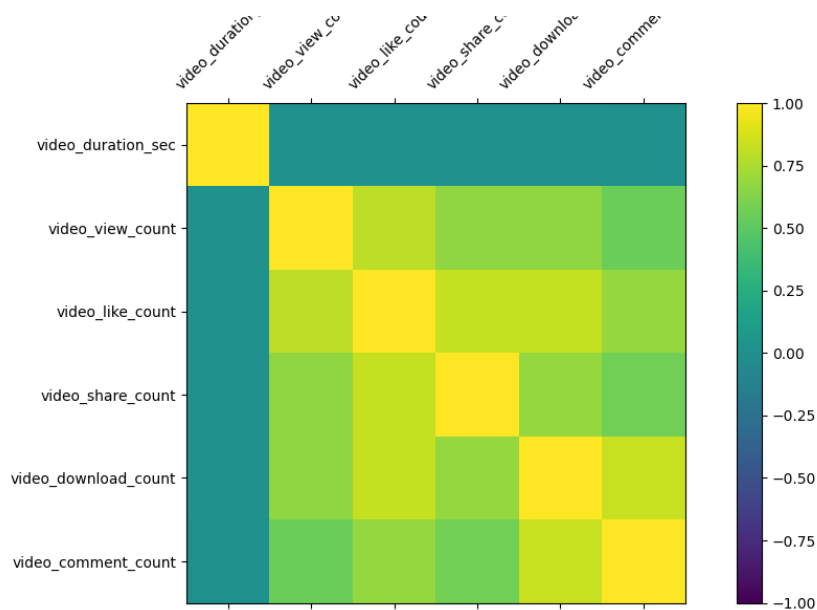
È stato scelto il seguente dataset: <https://www.kaggle.com/datasets/yakhyojon/tiktok>. Il dataset contiene circa 19mila entry riguardo a video pubblicati su TikTok, divisi in "affermazioni" e "opinioni". Con opinione si intende un pensiero di un singolo o di un gruppo, mentre un'affermazione è un'informazione 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



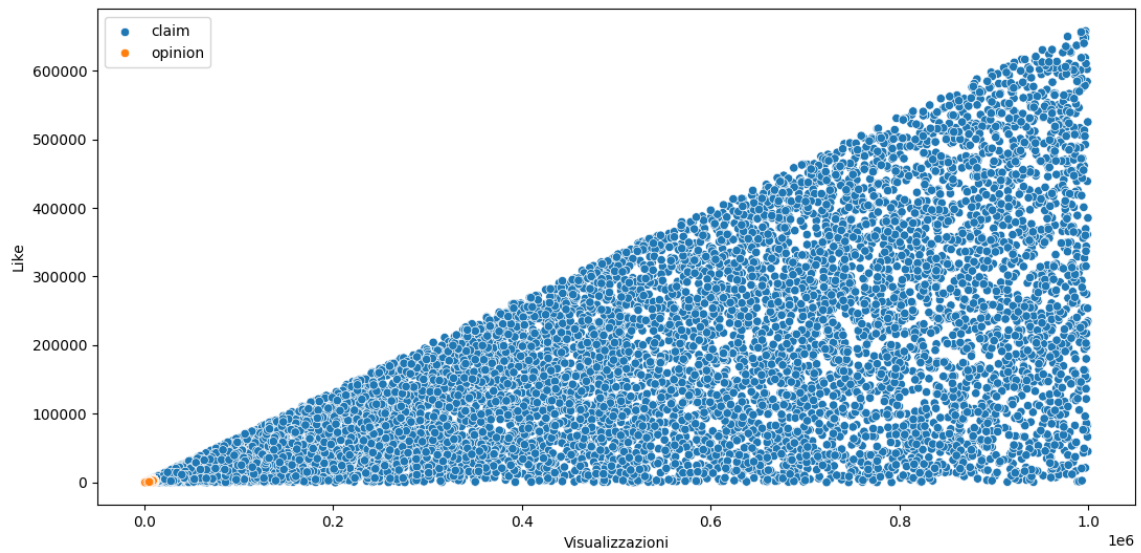
Matrice di correlazione tra le variabili numeriche del dataset.

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

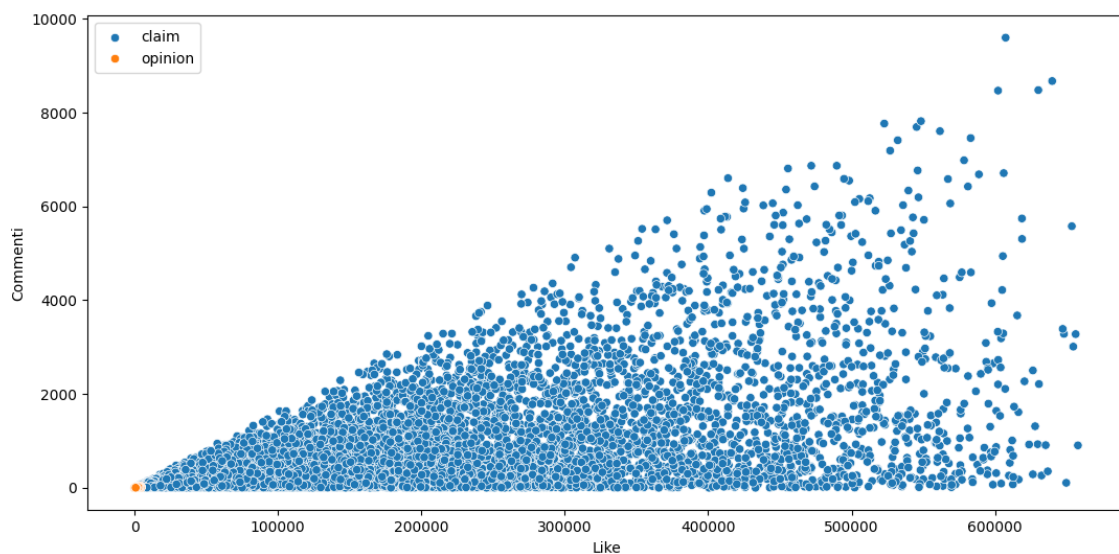
Le visualizzazioni hanno una correlazione positiva con le altre metriche: questo indica che il numero di visualizzazioni influenza fortemente le interazioni con il video (like, commenti,...). Lo stesso vale per il numero di commenti, condivisioni, download e commenti.

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 loro numero nel dataset sia quasi lo stesso.



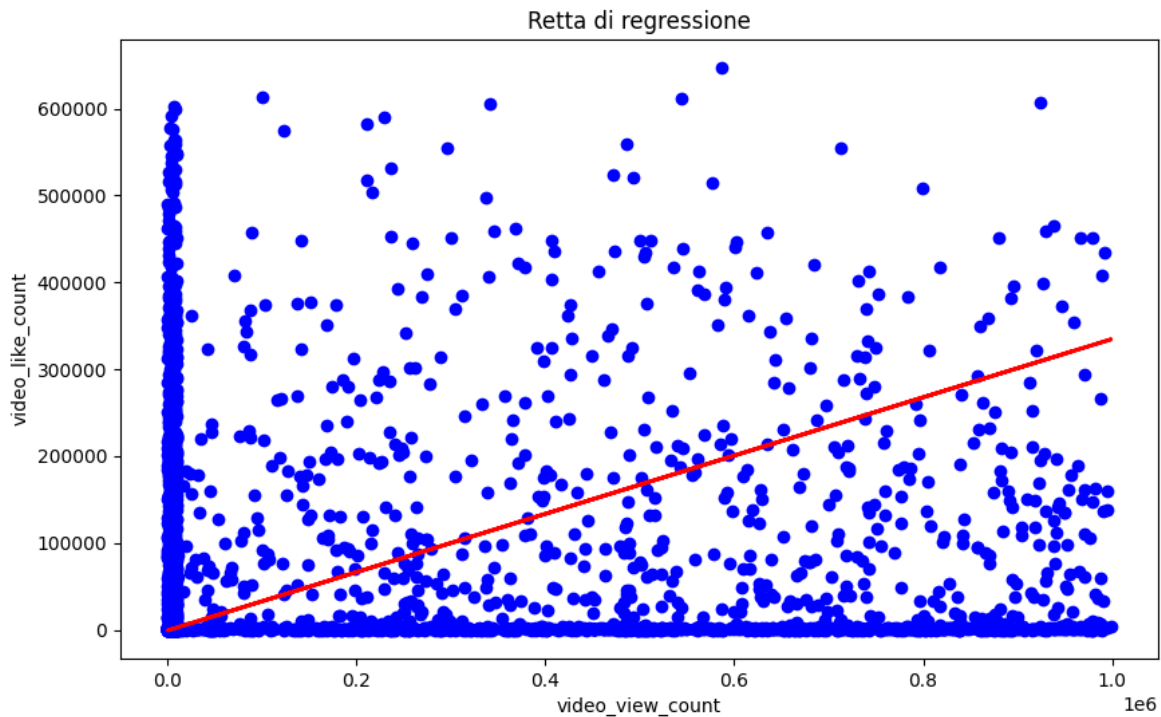
In questo grafico si nota che la densità dei punti diminuisce spostandosi verso destra e verso l'alto, indicando che i video con un altissimo numero di like e commenti sono meno comuni (o con numero di commenti paragonabile a quello delle visualizzazioni). Tuttavia anche in questo caso è presente una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

4. Splitting

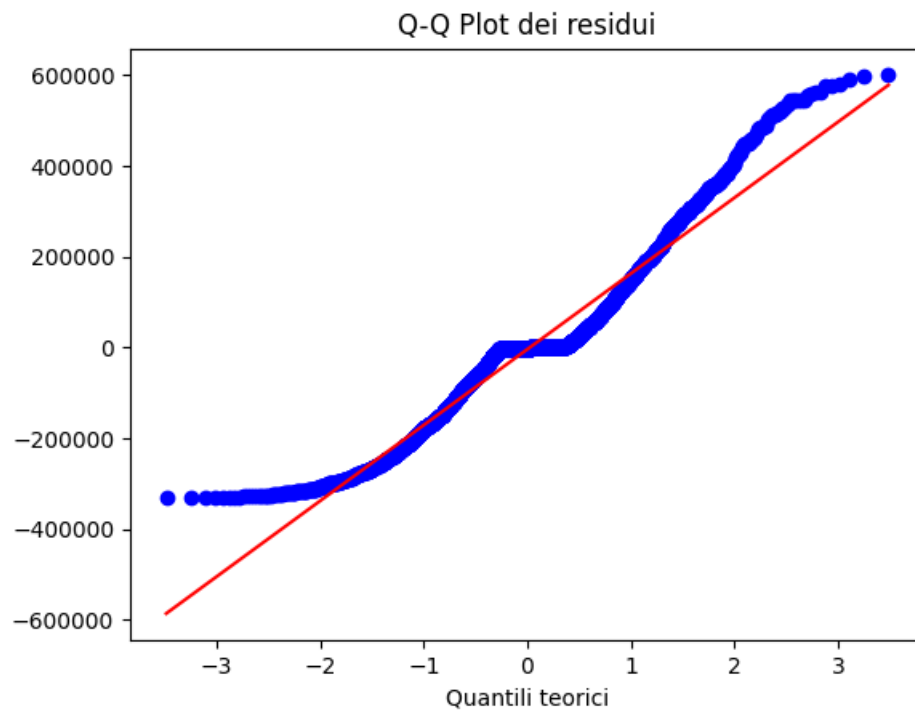
Il dataset è stato diviso in training set (70%), test set (15%) e validation set (15%).

5. Regressione Lineare

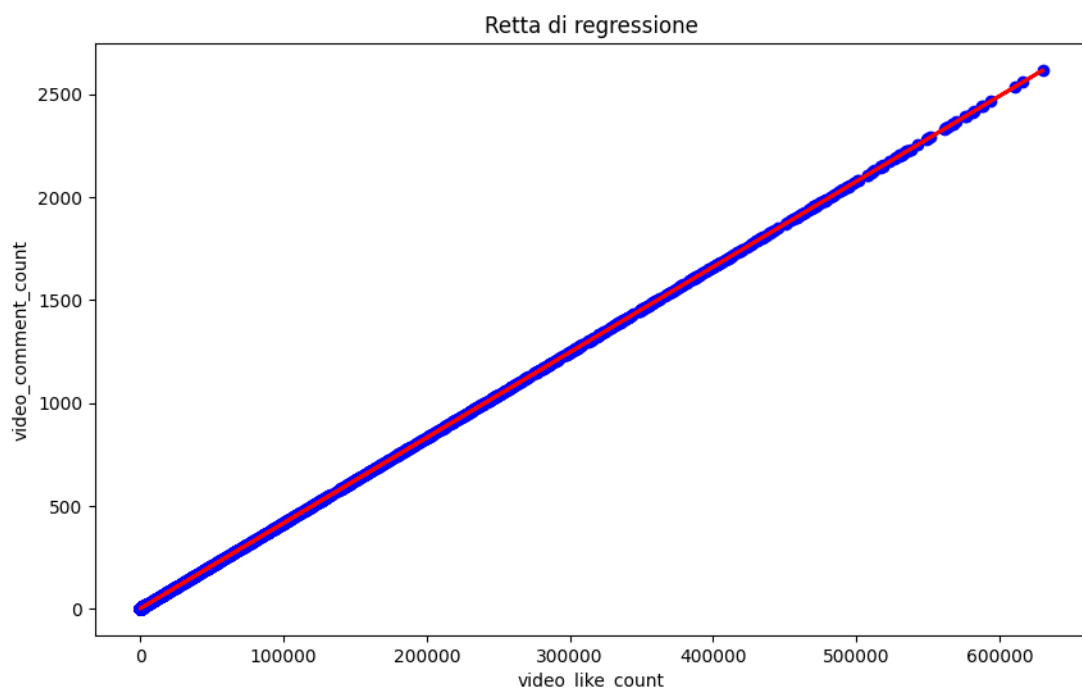
Si è eseguita la regressione lineare con $X=\text{video_view_count}$ e $y=\text{video_like_count}$ (fortemente correlati).



- Coefficiente angolare: 0.0041461075784415295. Questo valore indica la pendenza della retta di regressione.
- Intercetta: 2.98502114940203. 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 (ci si aspetta che video con 0 visualizzazioni non possano avere like).
- Coefficiente R^2 : -0.5507904468192637 . R^2 è stato calcolato con la funzione `r2_score` di `sklearn`, usando come input le y predette e le y del validation set. Il fatto che $R^2 \notin [0, 1]$ indica che la regressione lineare non è adatta a fare previsioni corrette su questo dataset. Per migliorare l'accuratezza del modello sono necessarie altre variabili in input, come vedremo con la regressione logistica e SVC.
- Errore quadratico medio (MSE): 896603.745500345. Indica che c'è una grande discrepanza tra i valori predetti dal modello e i valori effettivi.
- p-value: $2.6692226197466672 \times 10^{-48}$. C'è una probabilità molto bassa, praticamente nulla, che i dati siano distribuiti normalmente, quindi si rifiuta l'ipotesi di normalità H_0 .

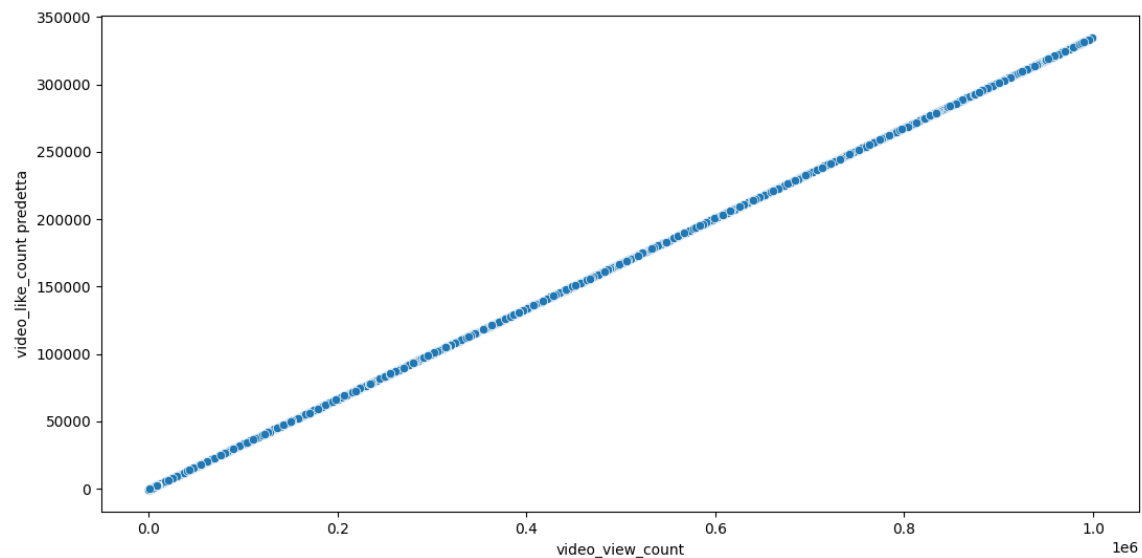


Il Q-Q plot indica in particolare che ci sono residui estremamente grandi più ci si allontana dal centro dei dati. Eseguendo la regressione tra like e commenti, anch'essi correlati positivamente, si ottengono sempre risultati non accurati.



6. Addestramento del modello

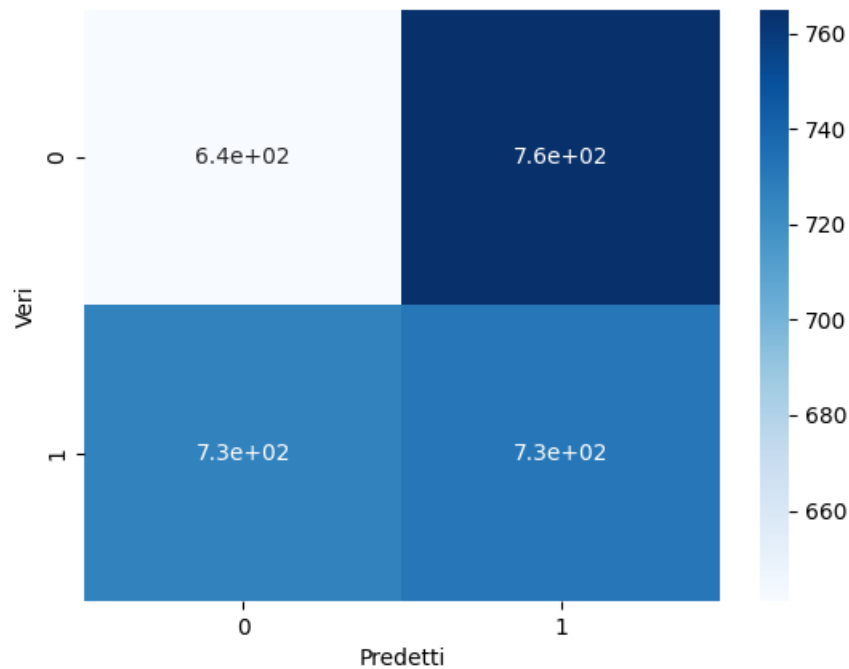
Si addestra il modello con regressione logistica per predire, dato il numero di visualizzazioni, like, commenti e condivisioni se un video è un'affermazione o un'opinione. Le X sono state standardizzate per far convergere l'algoritmo e migliorare la performance del modello.



Il report di classificazione afferma che c'è una precisione del 47% nel predire affermazioni, e del 49% per quanto riguarda le opinioni. Il 48% Delle predizioni totali sono risultate corrette.

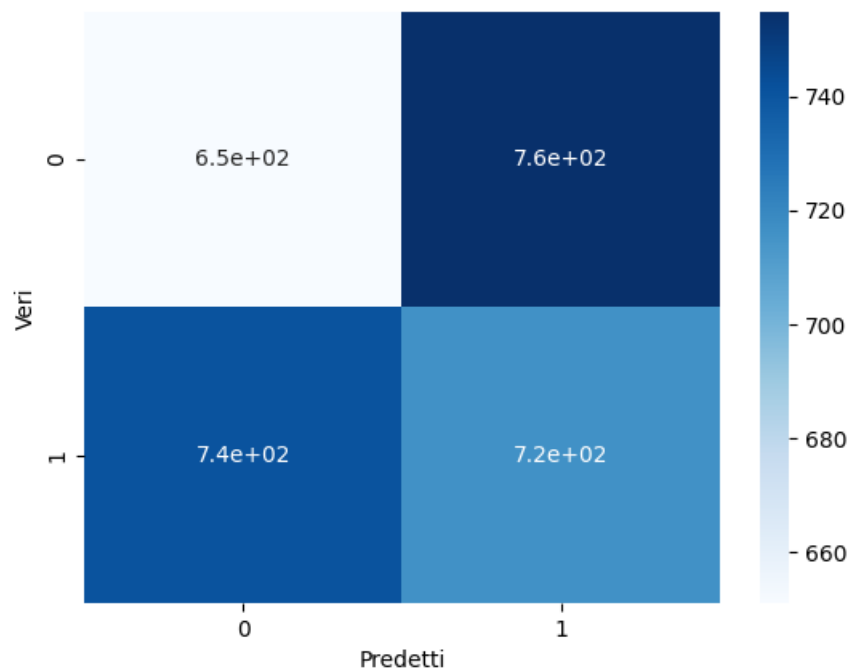
Guardando la matrice di confusione si nota che:

- 641 è il numero di veri negativi per la classe affermazione.
- 765 è il numero di falsi positivi per la classe affermazione.
- 726 è il numero di falsi negativi per la classe opinione.
- 731 è il numero di veri positivi per la classe opinione.



Con la Support Vector Classification (SVC) lineare, i risultati sono pressoché identici. Le precisioni nell'individuare le opinioni e le affermazioni rimane la stessa, e i valori nella matrice di confusione variano solo di

qualche decina



7. Hyperparameter Tuning

Per trovare il grado con accuratezza migliore in una SVC poly si eseguono più fasi di addestramento, con grado tra 1 e 5, tenendo traccia delle varie precisioni. Si nota che il grado che produce il risultato migliore è il quarto, anch'esso con precisione complessiva del 49%

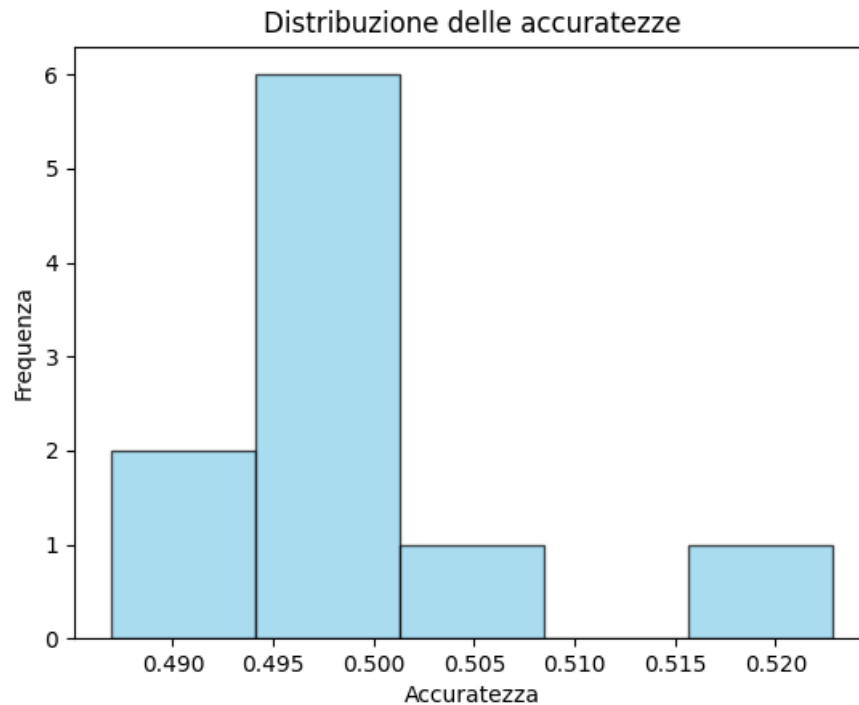
8. Valutazione della performance

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 4 e regressione logistica. SVC poly con grado dal primo al quinto ha le seguenti accuratezze: 0.479217603911, 0.485504715333, 0.479217603911, 0.487251135172, 0.477820468040.

9. Studio statistico sui risultati della valutazione

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

| Iterazione | Errore di classificazione (ME) | Tasso di errore di classificazione (MR) | Accuratezza (ACC) |
|------------|--------------------------------|---|---------------------|
| 1 | 1450 | 0.5064617534055187 | 0.4935382465944813 |
| 2 | 1440 | 0.5029689137268599 | 0.4970310862731401 |
| 3 | 1433 | 0.5005239259517988 | 0.49947607404820116 |
| 4 | 1438 | 0.5022703457911282 | 0.49772965420887183 |
| 5 | 1417 | 0.4949353824659448 | 0.5050646175340552 |
| 6 | 1442 | 0.5036674816625917 | 0.4963325183374083 |
| 7 | 1366 | 0.47712190010478517 | 0.5228780998952148 |
| 8 | 1445 | 0.5047153335661894 | 0.49528466643381064 |
| 9 | 1432 | 0.5001746419839329 | 0.4998253580160671 |
| 10 | 1469 | 0.5130981487949703 | 0.4869018512050297 |



L'accuratezza media è 0.49940621725462797, che è molto simile alla mediana: 0.49738037024100595. Questo suggerisce che le accuratezze non sono influenzate da valori estremi (perché la media è molto sensibile agli outlier e la mediana no). Una deviazione standard dello 0.08% 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 ha la forma di una campana gaussiana, si può affermare che la distribuzione dei residui non è normale.

Questi valori sono ritrovabili nel boxplot, che con la linea arancione indica la mediana. La base e altezza del rettangolo nel rappresentano gli estremi dell'intervallo interquartile (i valori maggiori del 25% e minori del 75% dei valori totali). I baffi indicano l'intervallo complessivo dei dati, al di fuori dei quali sono presenti due outlier (valori estremi) che tuttavia non sembrano influenzare molto la media, forse a causa del segno opposto.

L'intervallo di confidenza con $\alpha = 0.05$: [0.4938321730836435, 0.5049802614256125] 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 più bassi rispetto alle affermazioni, tutti i modelli sono in grado di predire correttamente quasi la metà delle volte la classe. Per quanto riguarda la regressione lineare, dato che un solo dato in input non è sufficiente per fare predizioni corrette sull'andamento delle visualizzazioni in base al numero di like, ci si ritrova con dati che non seguono assolutamente l'andamento atteso. Un numero maggiore di dati in addestramento potrebbe portare ad un'aumento dell'accuratezza da parte di tutti i modelli.