Francesco Forlani

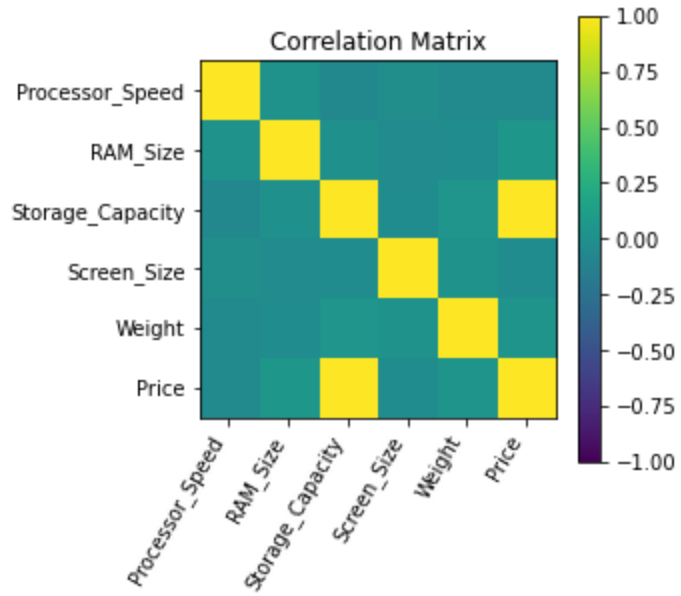
Statistica numerica

Discussione risultati progetto

Di seguito vengono commentati i risultati ottenuti nella parte descrittiva (EDA) e nella parte predittiva (regressione e classificazione). Nel progetto ho utilizzato un dataset preso da Kaggle, il quale contiene dati su prezzo, brand e specifiche di un insieme di laptops.

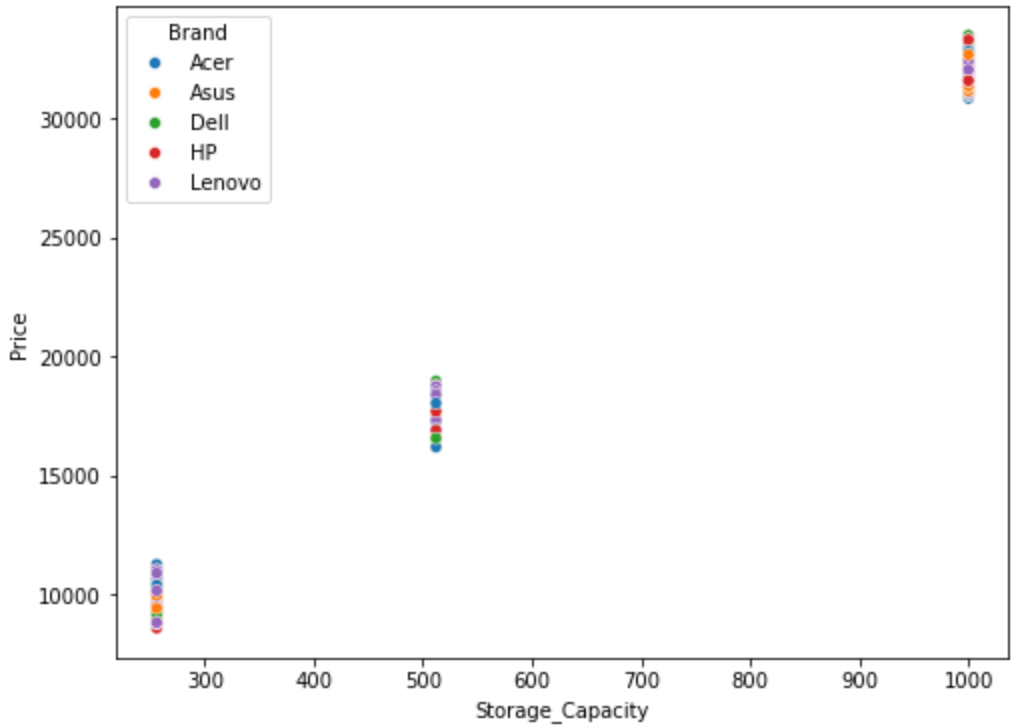
PARTE DESCRITTIVA (EDA)

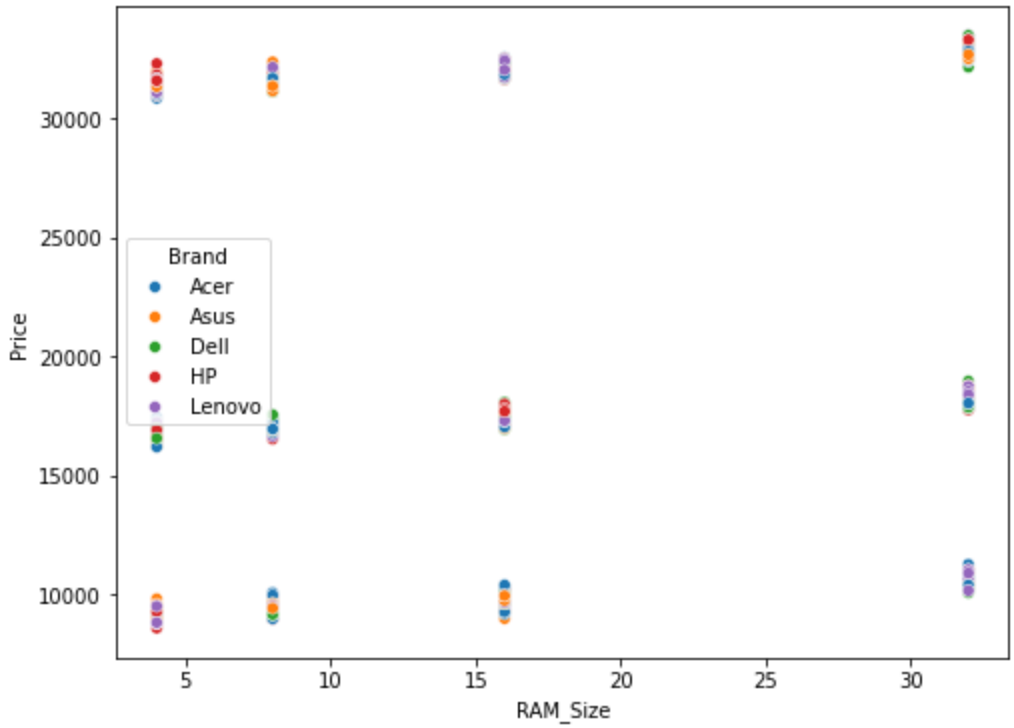
Nel codice ho utilizzato la matrice di Pearson per avere coscienza di come le variabili fossero correlate tra di loro.

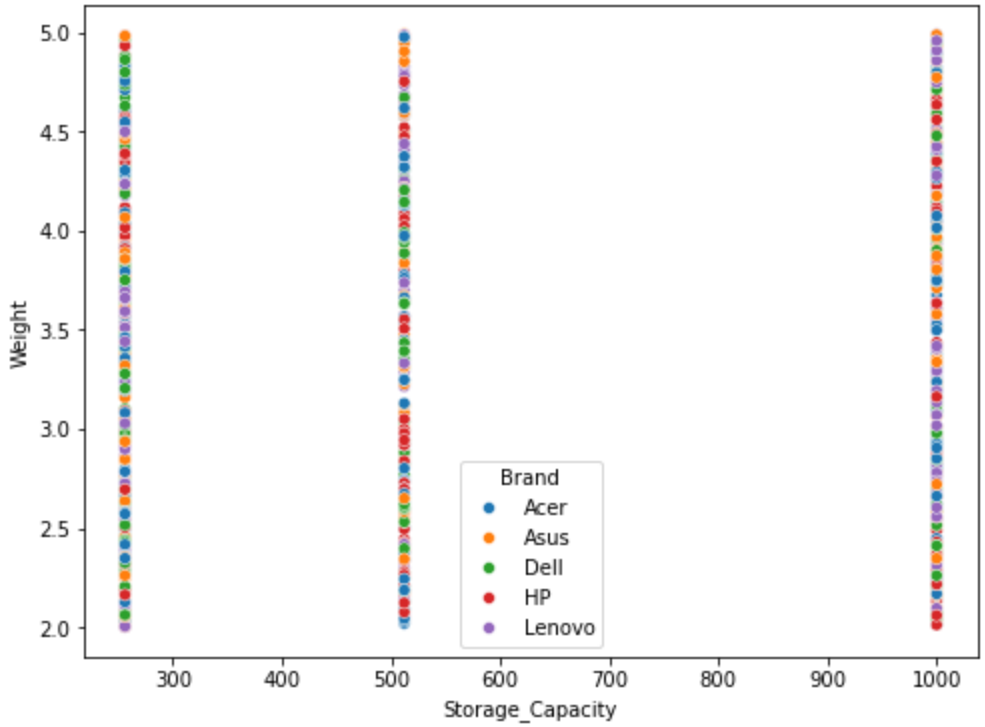


Da questa notiamo che ‘Storage Capacity’ e ‘Price’ hanno tra loro un indice di correlazione molto vicino a 1, 0.997908, il che indica una forte correlazione positiva, che significa che all'aumentare di una variabile, aumenta anche l'altra. Si nota inoltre che, tolta la precedente

coppia di variabili, tutte le altre hanno un indice di correlazione vicino allo 0, dunque ho optato per visualizzare i grafici delle 3 coppie con indice più elevato in modo da poter notare al meglio le relazioni più rilevanti. Oltre la precedente coppia, le altre due scelte sono ‘Price’ e ‘RAM Size’ ; ‘Storage capacity’ e ‘Weight’ .




Price
RAM_Size
Brand
Acer
Asus
Dell
HP
Lenovo
30000
25000
20000
15000
10000
5
10
15
20
25
30


Weight
Storage_Capacity
Brand
Acer
Asus
Dell
HP
Lenovo
5.0
4.5
4.0
3.5
3.0
2.5
2.0
300
400
500
600
700
800
900
1000

Dai grafici a dispersione utilizzati, notiamo che, ‘Storage Capacity’ e ‘Price’ , confermano la loro forte relazione lineare positiva, dunque la capacità di archiviazione è il principale fattore che fa aumentare il prezzo di un laptop. Tuttavia, la presenza di cluster distinti nel grafico suggerisce la presenza di altri fattori che contribuiscono a differenziare i prezzi tra dispositivi con capacità di archiviazione simile. Il più influente di questi altri fattori è infatti la dimensione della RAM. Analizzando la coppia ‘Price’ e ‘RAM Size’, notiamo dal grafico che, nonostante l’indice di correlazione non sia elevato (0.061237), all’aumentare della RAM aumenta anche il prezzo, sebbene molto più lentamente rispetto alla capacità di archiviazione. Anche in questo grafico, cosa comune a tutti quelli presi in considerazione, i dati sono raggruppati in clusters, cosa dovuta al fatto che l’archiviazione e la RAM possono assumere solo certe dimensioni (256,512,... ; 4, 8 ,16,... GB).

Il grafico dell’ultima coppia presa in considerazione, la terza per indice di correlazione (0.041335) , ‘Storage capacity’ e ‘Weight’, mostra che, seppur molto debolmente, all’aumentare dello spazio di archiviazione, aumenta anche il peso. Tuttavia, i dati sono raggruppati in 3 clusters verticali che differiscono leggermente nell’intervallo di peso in cui si trovano. Ciò indica che la capacità di archiviazione è solo uno dei tanti fattori che incide sul peso di un dispositivo.

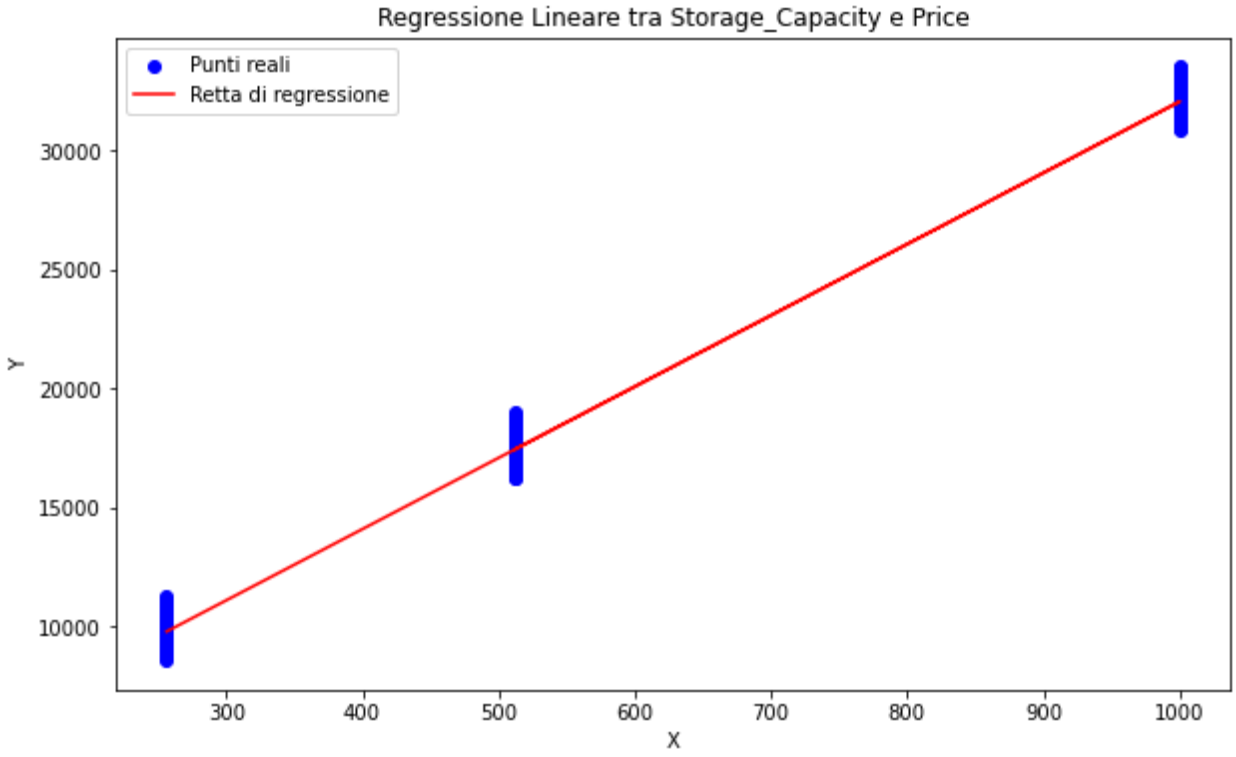
CONCLUSIONE EDA

Il prezzo di un dispositivo è influenzato in maniera determinante dalla capacità di archiviazione. Ci sono altri fattori, come la dimensione della RAM, che lo influenzano ma in maniera molto meno impattante. In generale, tutte le altre coppie di fattori, hanno un indice di correlazione molto vicino allo 0, il che indica che il mutare di uno influenza poco gli altri fattori.

PARTE PREDITTIVA (REGRESSIONE E CLASSIFICAZIONE)

Regressione lineare

La regressione lineare è stata effettuata tra le due coppie con indice di correlazione più elevato, che, come scritto precedentemente, sono le coppie 'Storage Capacity' e 'Price' ; 'RAM_Size' e 'Price'.



Per la prima coppia sono stati ottenuti i seguenti risultati:

- **Coefficiente di Determinazione (R^2):** 0,9958
- **Mean Squared Error (MSE):** 369445,48
- **Root Mean Squared Error (RMSE):** 607,82
- **Percentuale rispetto alla media della variabile dipendente:** 3,10%

- **Test di Shapiro-Wilk per la normalità dei residui:**
    - **Statistic:** 0,9534
    - **p-value:** 2,84e-17

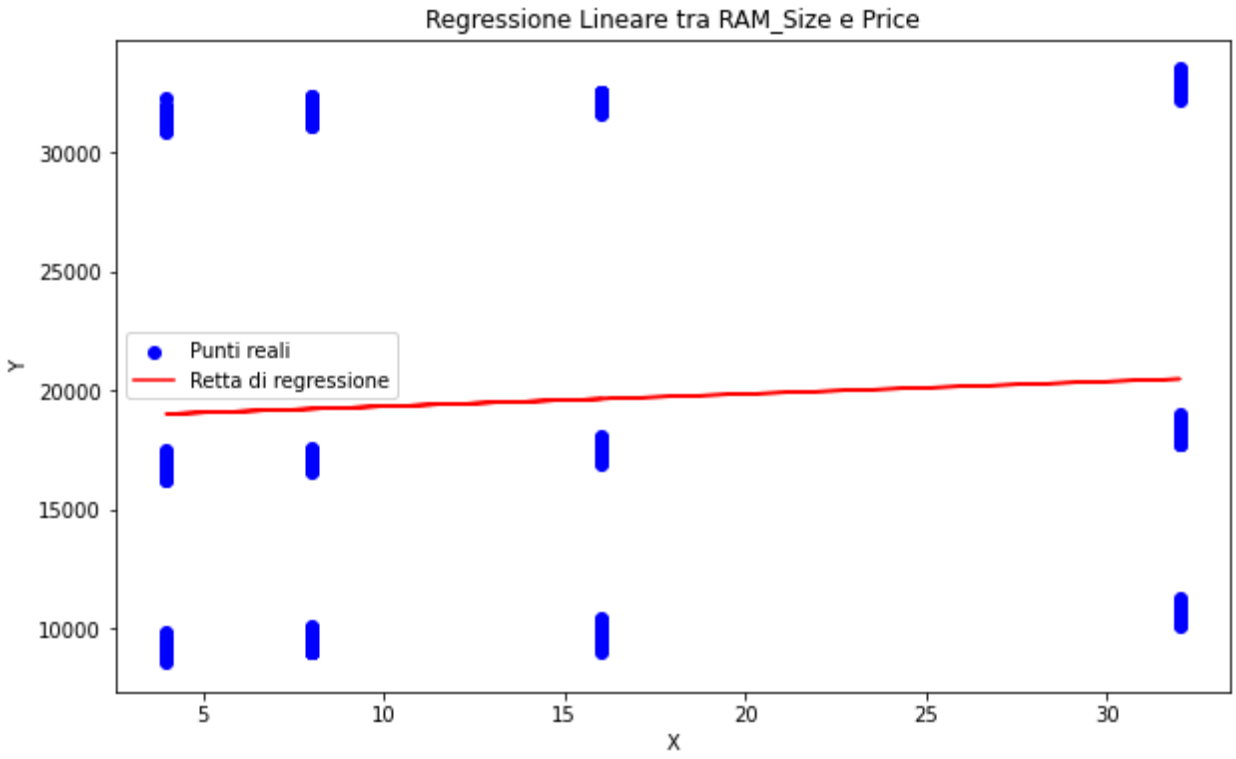
Il coefficiente di determinazione (R^2) di 0,9958 indica che il modello spiega il 99,58% della variabilità della variabile dipendente, suggerendo un'ottima capacità predittiva. Tuttavia, il valore del MSE è piuttosto elevato, con un RMSE di 607,82, il che indica una discreta dispersione dei valori predetti rispetto ai valori osservati. Questo errore rappresenta circa il 3,1% della media della variabile dipendente, suggerendo che, pur essendo il modello preciso nel spiegare la variabilità, gli errori predittivi sono comunque significativi in termini assoluti.

Il test di Shapiro-Wilk per la normalità dei residui restituisce un p-value estremamente basso (2,84e-17), suggerendo che i residui non seguono una distribuzione normale. Questo è probabilmente indice di una specifica struttura nei dati che il modello lineare non riesce a catturare adeguatamente.

Per quanto riguarda la seconda coppia sono stati ottenuti i seguenti dati:

- **Coefficiente di Determinazione (R^2):** 0,0037
- **Mean Squared Error (MSE):** 88054142
- **Root Mean Squared Error (RMSE):** 9383,72
- **Percentuale rispetto alla media della variabile dipendente:** 47,86
- **Test di Shapiro-Wilk per la normalità dei residui:**
    - **Statistic:** 0,78

- **p-value:** 1.48e-34



Il secondo modello presenta un coefficiente di determinazione (R^2) molto basso, pari a 0,0037, il che indica che il modello non riesce a spiegare in modo significativo la variabilità della variabile dipendente. Il valore del MSE è notevolmente superiore rispetto al primo modello, con un RMSE di 9383,72 . La percentuale di errore rispetto alla media della variabile dipendente è molto alta (47,86%), dunque, il modello fa predizioni con un errore medio troppo elevato rispetto al prezzo medio, rendendo i risultati poco utili.. Il test di Shapiro-Wilk mostra ancora una volta un p-value estremamente basso (1.48e-34), confermando che i residui non sono normalmente distribuiti.

CONCLUSIONE REGRESSIONE LINEARE:

Il primo modello mostra un'elevata capacità predittiva in termini di R^2, ma con un errore predittivo ancora rilevante. Il secondo modello, al contrario, non riesce a spiegare in modo significativo la variabilità della variabile dipendente. In entrambi i modelli, i residui non seguono una distribuzione normale. Tuttavia, è da ricordare che il modello di regressione lineare semplice è robusto anche sufficientemente lontano dalla normalità dei residui, dunque il primo modello è ragionevolmente valido e si può in pratica utilizzare.

Classificazione

La variabile target che il modello addestrato cerca di predire è se il prezzo di un laptop è alto o basso, identificandolo come alto se maggiore della mediana, basso viceversa.

Per quanto riguarda la regressione logistica, i risultati ottenuti sono stati i seguenti:

Accuratezza del Modello

Il modello ha raggiunto un'accuratezza del 95,5% sul set di validazione. Questo significa che il 95,5% delle osservazioni è stato classificato correttamente dal modello. Un'accuratezza così elevata indica che il modello è molto efficace nel compito di classificazione, riuscendo a distinguere correttamente tra le classi con un margine di errore minimo.

Matrice di Confusione

[[100 0]

[9 91]]

- **Classe 0 (Negativa):** Il modello ha classificato correttamente tutte le 100 osservazioni della classe 0, senza alcun falso positivo.

- **Classe 1 (Positiva):** Il modello ha classificato correttamente 91 delle 100 osservazioni della classe 1, con 9 esempi classificati erroneamente come classe 0 (falsi negativi).

Report di Classificazione

- **Precisione, Recall e F1-Score:**
  - **Classe 0:**
    - **Precisione:** 0,92 (92%) - Il 92% delle osservazioni classificate come "0" appartengono effettivamente alla classe 0.
    - **Recall:** 1,00 (100%) - Tutte le osservazioni della classe 0 sono state correttamente identificate.
    - **F1-Score:** 0,96 (96%) - Unisce precisione e recall, indicando un'ottima performance nella classificazione della classe 0.
  - **Classe 1:**
    - **Precisione:** 1,00 (100%) - Tutte le osservazioni classificate come "1" appartengono effettivamente alla classe 1.
    - **Recall:** 0,91 (91%) - Il modello ha identificato correttamente il 91% delle osservazioni della classe 1, con un 9% di falsi negativi.
    - **F1-Score:** 0,95 (95%) - Indica un buon equilibrio tra precisione e recall nella classificazione della classe 1.

Il modello di regressione logistica ha mostrato prestazioni quasi perfette nel set di validazione, con un'accuratezza del 95,5% e ottimi risultati in termini di precisione, recall e F1-score per

entrambe le classi. Tuttavia, è emerso un lieve margine di miglioramento nel riconoscimento della classe 1, dove sono stati osservati alcuni falsi negativi.

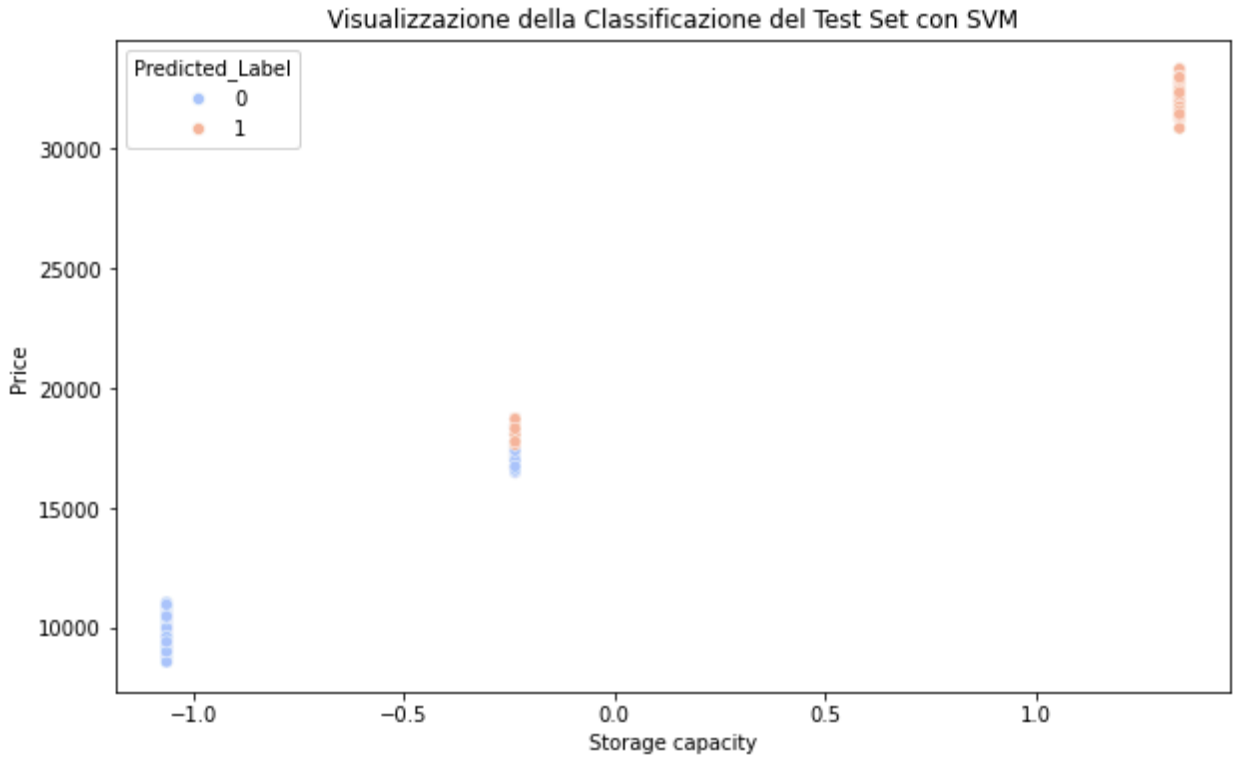
Per quanto riguarda SVC, i parametri migliori sono stati determinati basandosi su una ricerca esaustiva (GridSearchCV) che ha esplorato diverse combinazioni possibili di iperparametri. La combinazione che ha fornito la migliore accuratezza durante la cross-validation è stata selezionata come la migliore, e questo modello è stato poi valutato sui dati di validazione per verificare la sua capacità di generalizzare su dati nuovi.

I migliori iperparametri che sono stati selezionati sono: **{'C': 1, 'gamma': 'scale', 'kernel': 'linear'}**. I risultati sul validation set sono stati:

- **Accuracy:** 0.97
- **Matrice di confusione:**
    - True Positives (TP): 98, True Negatives (TN): 96
    - False Positives (FP): 2, False Negatives (FN): 4
- **Classification Report:**
    - Precisione: 0.96 per la classe 0 e 0.98 per la classe 1.
    - Recall: 0.98 per la classe 0 e 0.96 per la classe 1.
    - F1-Score: 0.97 per entrambe le classi.

Questi risultati indicano che il modello funziona bene su entrambe le classi senza favorire una rispetto all'altra. Avendo con questo modello ottenuto i risultati migliori in assoluto, ho proceduto alla valutazione delle performance. I risultati sul test set sono stati:

- **Accuracy:** 0.965
- **Matrice di confusione:**
    - TP: 100, TN: 93
    - FP: 0, FN: 7
- **Classification Report:**
    - Precisione: 0.93 per la classe 0 e 1.00 per la classe 1.
    - Recall: 1.00 per la classe 0 e 0.93 per la classe 1.
    - F1-Score: 0.97 per la classe 0 e 0.96 per la classe 1.



Sul test set, il modello mostra una leggera diminuzione della recall per la classe 1 (93% rispetto al 96% sul validation set), ma una precisione perfetta per la stessa classe. Questo suggerisce che,

pur mantenendo alta la capacità di evitare falsi positivi, il modello può aver perso alcuni veri positivi per la classe 1.

Successivamente, ho ripetuto le fasi di addestramento e testing 10 volte per questo modello, in modo da ottenere un SRS(10) di ogni metrica, i risultati sono stati:

- **MSE:**
    - Media: 0.0235, con un intervallo di confidenza al 95% tra 0.0172 e 0.0298.
    - L'MSE basso indica che l'errore medio è contenuto.
- **Accuracy:**
    - Media: 0.9765, con un intervallo di confidenza al 95% tra 0.9702 e 0.9828.
    - L'accuratezza è molto alta e costante tra le iterazioni.
- **Precision:**
    - Media: 0.9708, con un intervallo di confidenza al 95% tra 0.9589 e 0.9826.
    - La precisione ha una variabilità relativamente bassa, con un valore minimo di 0.9429, mostrando un'alta capacità del modello di evitare falsi positivi.
- **Recall:**
    - Media: 0.9830, con un intervallo di confidenza al 95% tra 0.9713 e 0.9947.
    - La recall ha il valore medio più alto tra le metriche, indicando una bassa probabilità di falsi negativi.
- **F1 Score:**
    - Media: 0.9767, con un intervallo di confidenza al 95% tra 0.9704 e 0.9829.

- L'F1 score conferma un buon equilibrio tra precisione e recall, con una bassa variabilità.

CONCLUSIONE CLASSIFICAZIONE

Il modello SVM presenta ottime performance sia sul validation set che sul test set, con metriche molto simili tra loro. Le iterazioni successive dell'addestramento mostrano che il modello è robusto e consistente, con variazioni minime nelle metriche di errore. L'accuratezza media è elevata (97.65%) e l'analisi statistica conferma la stabilità del modello con intervalli di confidenza stretti.

Questi risultati indicano che l' SVM è ben addestrato e generalizza bene sui dati non visti. Le prestazioni sono solide e le metriche non mostrano segni di overfitting o underfitting significativi.