Relazione Progetto Statistica

Scelta Dataset

Il dataset che ho deciso di utilizzare contiene 1000 righe e 8 colonne, con variabili sia categoriche che numeriche.

Spiegazione Colonne:

1. **gender**: Genere dello studente (maschio o femmina).
2. **race/ethnicity**: Appartenenza etnica.
3. **parental level of education**: Livello di istruzione dei genitori.
4. **lunch**: Tipo di pranzo (standard o gratuito/ridotto).
5. **test preparation course**: Se hanno completato un corso di preparazione al test.
6. **math score**: Punteggio nel test di matematica.
7. **reading score**: Punteggio nel test di lettura.
8. **writing score**: Punteggio nel test di scrittura.

Tutte le colonne sono complete, senza valori mancanti.

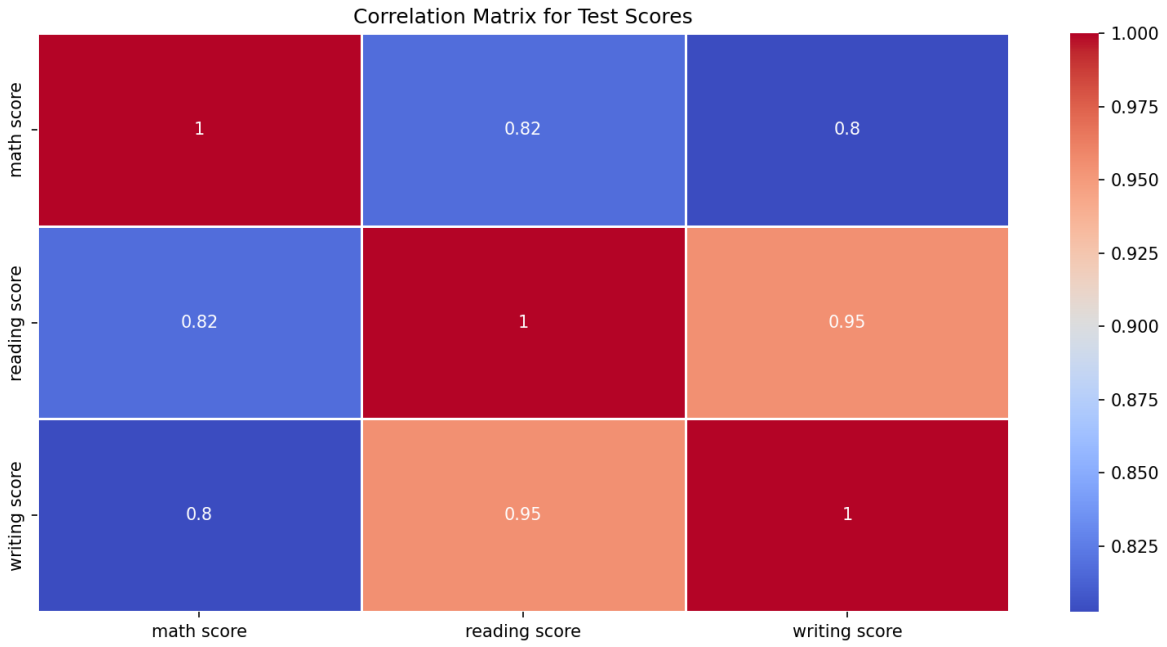
EDA

Matrice di correlazione

```
    # Matrice di correlazione per i punteggi dei test
    corr_matrix = data[['math score', 'reading score', 'writing

    # Heatmap della matrice di correlazione
```

```
    plt.figure(figsize=(8,6))
    sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewi
    plt.title('Correlation Matrix for Test Scores')
    plt.show()
```



La matrice di correlazione mostra una forte correlazione positiva tra i punteggi di lettura e scrittura (r=0.95)

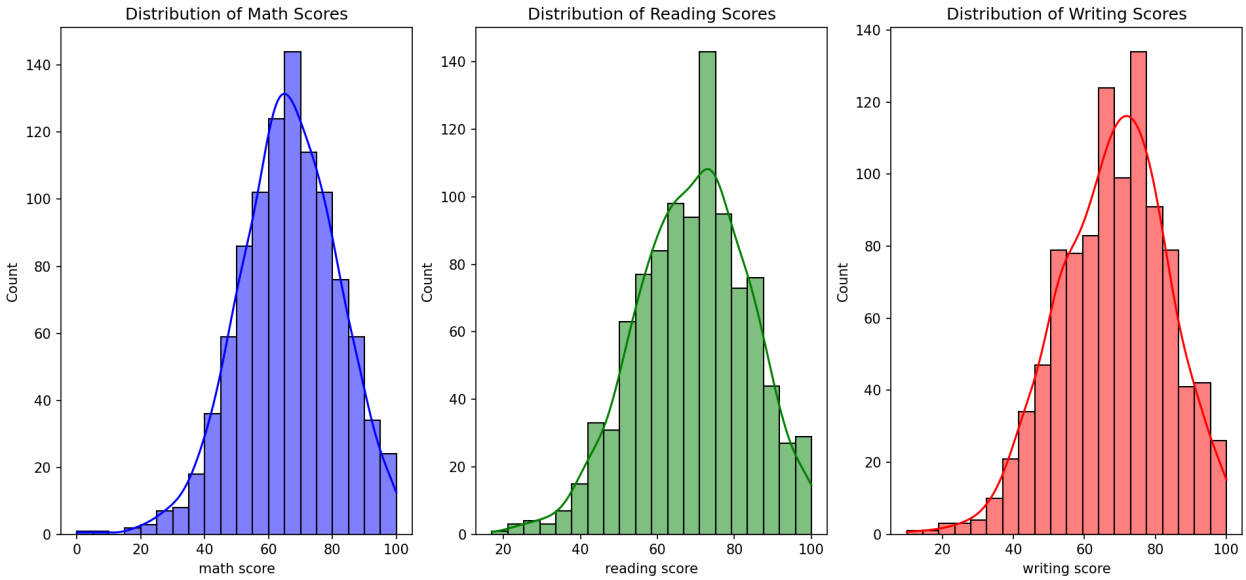
seguita da una correlazione più moderata tra lettura e matematica (r=0.55) e scrittura e matematica (r=0.58)

Distribuzione delle proprietà

```
# Distribuzione univariata dei punteggi dei test
    plt.figure(figsize=(14,6))

    # Istogramma per il punteggio di matematica
    plt.subplot(1, 3, 1)
```

```
    sns.histplot(data['math score'], bins=20, kde=True, color='b
    plt.title('Distribution of Math Scores')
```

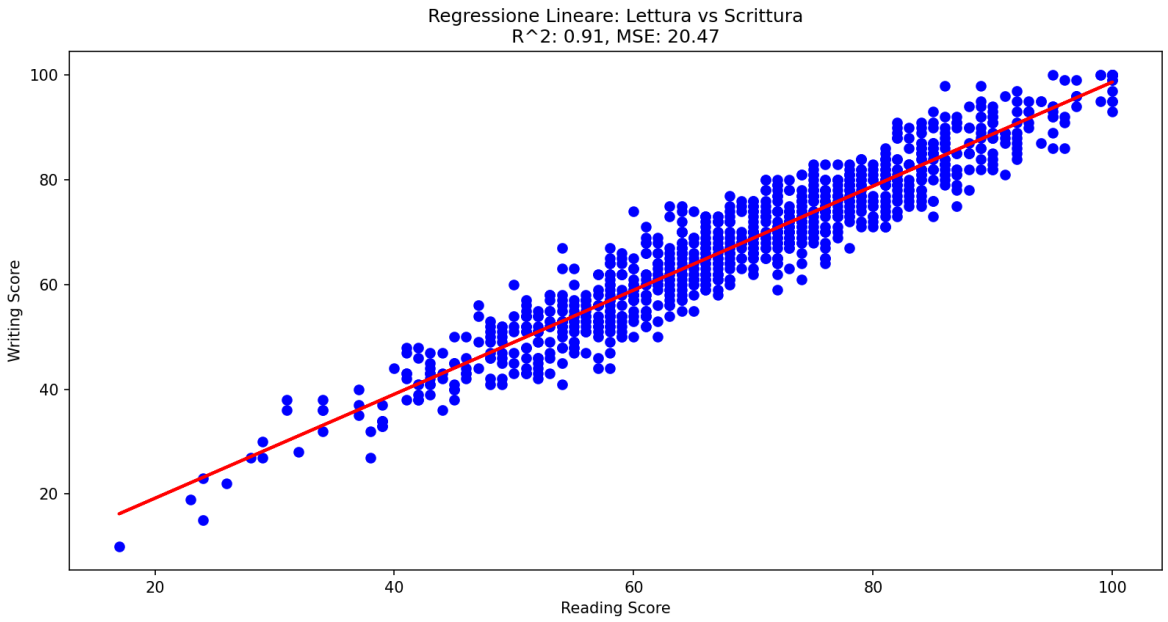


Gli istogrammi dei punteggi di matematica, lettura e scrittura mostrano una distribuzione abbastanza normale, con alcuni valori molto bassi ma nessun outlier evidente. Le distribuzioni sono leggermente sbilanciate verso il basso, specialmente per la matematica.

Parte Predittiva

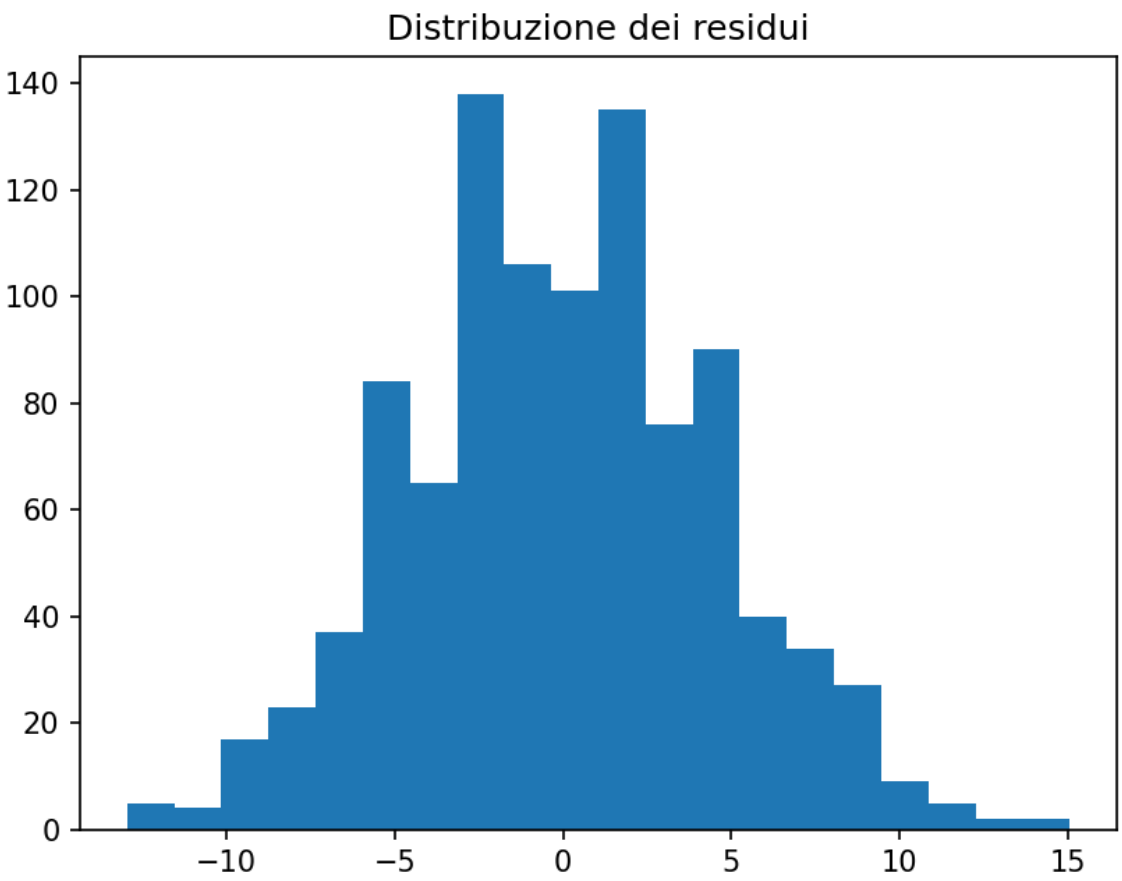
Regressione Lineare

Partendo dalle informazioni date dalla matrice di correlazione, eseguo una regressione lineare tra alcune delle variabili numeriche, nello specifico tra i punteggi di lettura e scrittura, che dimostrano di avere una forte correlazione.



Il coefficiente **R^2** ci dice quanto bene la linea di regressione approssima i dati osservati. Un valore di 0.91 indica che il 91% della varianza nei punteggi di scrittura può essere spiegato dai punteggi di lettura. Quindi valore di R^2 così alto suggerisce che il modello di regressione è molto efficace nel predire i punteggi di scrittura a partire dai punteggi di lettura.

L'MSE ottenuto è 20.47. Dato che i punteggi variano tra 10 e 100 e la deviazione standard è di circa 15.20 per i punteggi di scrittura, un MSE di 20.47 non è trascurabile, ma comunque rappresenta un buon risultato.



La
distribuzione dei residui sembra essere **simmetrica e** riporta una forma quasi vicina a quella di una distribuzione normale, quuesto è un buon segno per la validità del modello di regressione lineare, dato che uno degli assunti di base della regressione lineare è che i residui siano distribuiti normalmente intorno allo zero.
Il picco della distribuzione è vicino a 0, il che indica che la maggior parte delle differenze tra i valori predetti e quelli osservati è prossima allo zero.
Si osservano alcune code sia a sinistra che a destra (residui che si estendono da circa -10 a 15), questi casi però sono meno frequenti, indicando che ci sono alcuni outlier o dati in cui il modello non ha funzionato altrettanto bene.

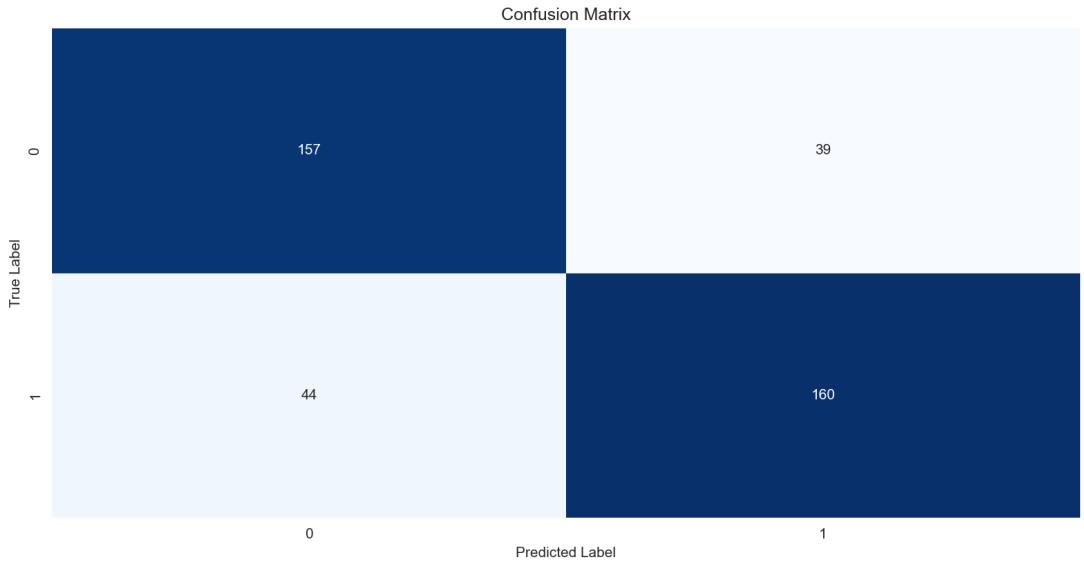
Regressione Logistica

Dato che il dataset non sembrava avere una variabile di classificazione immediata e con una buona correlazione con altre variabili, ho crearo una variabile target categorica basata, sul punteggio di matematica (dividendo gli studenti in due classi: "punteggio alto" e "punteggio basso", ovvero sopra o sotto la media). Quindi il modello, utilizzando le variabili del punteggio di scrittura e di lettura ha l'obiettivo di predirre se il punteggio di matematica è alto o basso.

Metriche

Accuratezza = 79.25% → L'accuratezza rappresenta la percentuale di campioni correttamente classificati. Quasi l'80% di accuratezza è ragionevolmente buona per un classificatore relativamente semplice come la regressione logistica.

Misclassification Rate MR = 20.75% →Rappresenta la percentuale di errori, ed è complementare all'accuratezza (100% - 79.25% = 20.75%). Questo valore indica che il 20.75% delle previsioni sono state errate.



La matrice di confusione risultante mostra:

1. Veri Positivi (TP): 160
2. Veri Negativi (TN): 157
3. Falsi Positivi (FP): 39
4. Falsi Negativi (FN): 44

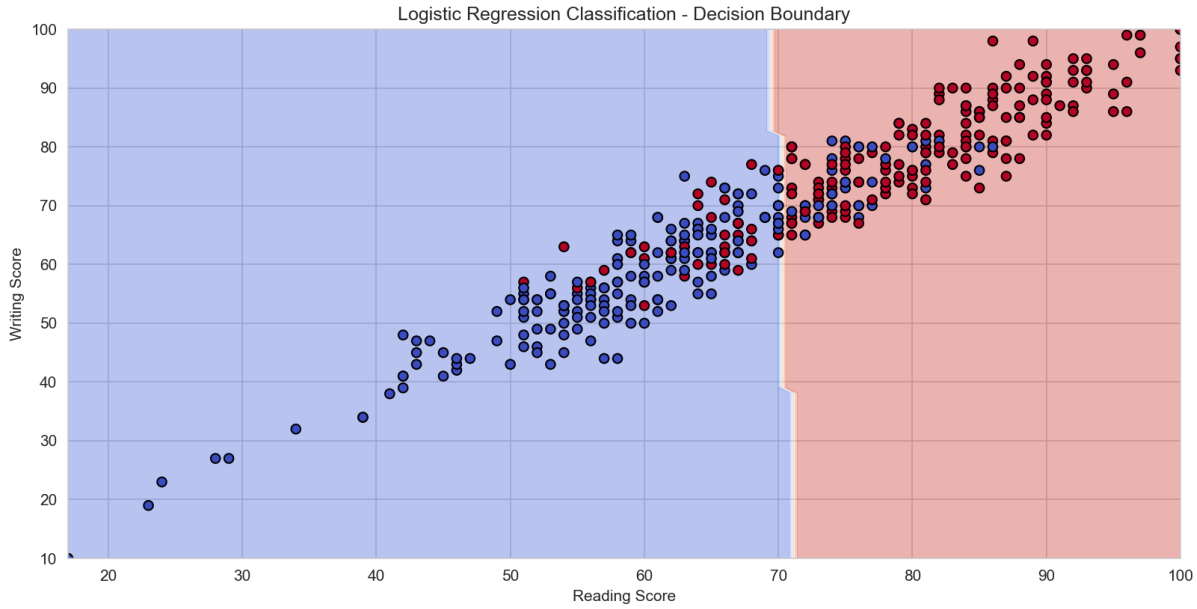
Da cui ricaviamo che:

Accuratezza: (TP + TN) / (TP + TN + FP + FN) = (160 + 157) / (160 + 157 + 39 + 44) =79.2%
Il modello classifica correttamente il 79.2% di tutti i casi, che è un buon risultato (già analizzato nella metrica precedente)

Precisione: TP / (TP + FP) = 160 / (160 + 39) =80.4%
Quando il modello predice un punteggio di matematica alto, ha ragione nell'80.4% dei casi.

Sensibilità: TP / (TP + FN) = 160 / (160 + 44) = 78.4%
Il modello identifica correttamente il 78.4% di tutti gli studenti con punteggi di matematica effettivamente alti.

Specificità: TN / (TN + FP) = 157 / (157 + 39) = 80.1%
Il modello identifica correttamente l'80.1% degli studenti con punteggi di matematica bassi.



Il grafico del decision boundary (la linea che separa la decisione) è cosi strutturato

- I punti blu rappresentano gli studenti con punteggi di matematica bassi
- I punti rossi rappresentano gli studenti con punteggi di matematica alti
- L'area blu chiaro rappresenta la regione dove il modello prevede punteggi di matematica bassi
- L'area rossa chiaro rappresenta la regione dove il modello prevede punteggi di matematica alti

Da cui si può evincere che Il modello riesce a separare abbastanza bene le due classi, ma ci sono alcune sovrapposizioni, specialmente nella zona centrale. Si possono notare alcuni punti blu nell'area rossa e viceversa, che rappresentano le classificazioni errate del modello (falsi positivi e falsi negativi).
Inoltre ci sono alcuni punti isolati, specialmente nella parte inferiore sinistra del grafico, che potrebbero rappresentare casi particolari o potenziali outlier.

SVM

Struggle

> Nel modello SVM ho dovuto implementare una tecnica di “scalatura” delle variabili, in quanto nel dataset di training erano presenti variabili con scala molto diversa, infatti avevo features con valori da 0 a 100 (ad esempio il punteggio di lettura) e un'altra con valori da 0 a 10 (ad esempio un punteggio di partecipazione).Questo problema di scala portava il modello, nel caso di utilizzo di kernel polinomiale, a non convergere mai e ha bloccare il processo di training.

Ho utilizzato la classe StandardScaler() di Sklearn

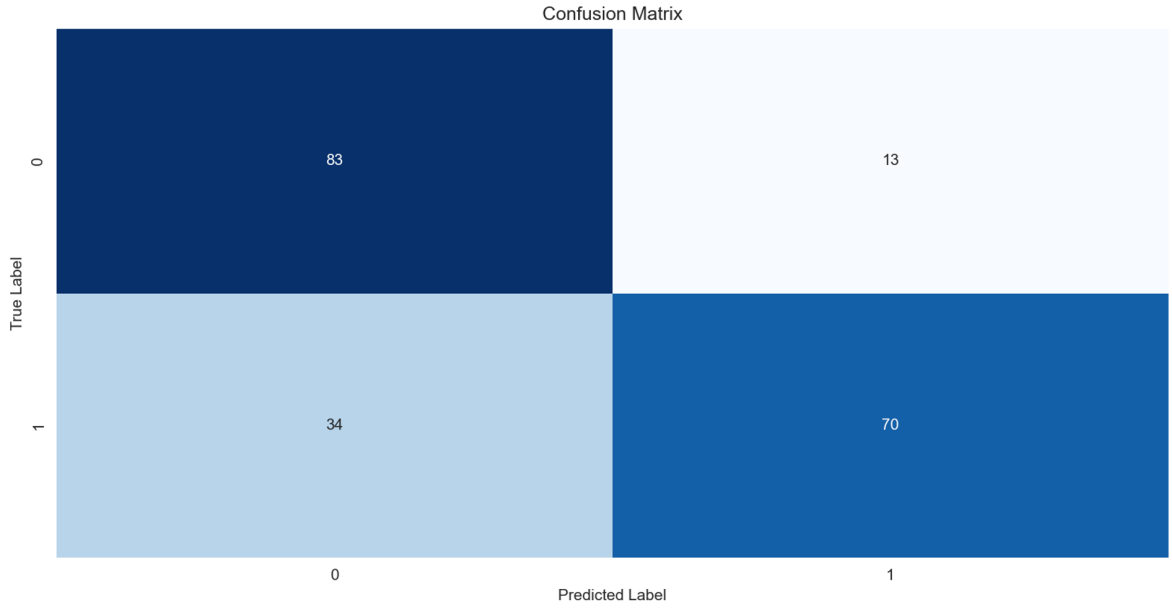
HyperParameter Tuning

Il risultato dell’utilizzo k-CV seleziona i seguenti parametri:

Miglior modello complessivo: SVM con kernel poly

Migliori parametri complessivi: {'C': 100, 'degree': 3, 'gamma': 'scale'}

Valutazione Modello



I risultati di questa validazione ci suggeriscono che l'accuratezza e la precisione sono buone, con valori rispettivamente del 76.50% e 84.34%, quindi significa che il modello ha predetto correttamente circa 3 predizioni su 4 e tende a non effettuare troppi falsi positivi .

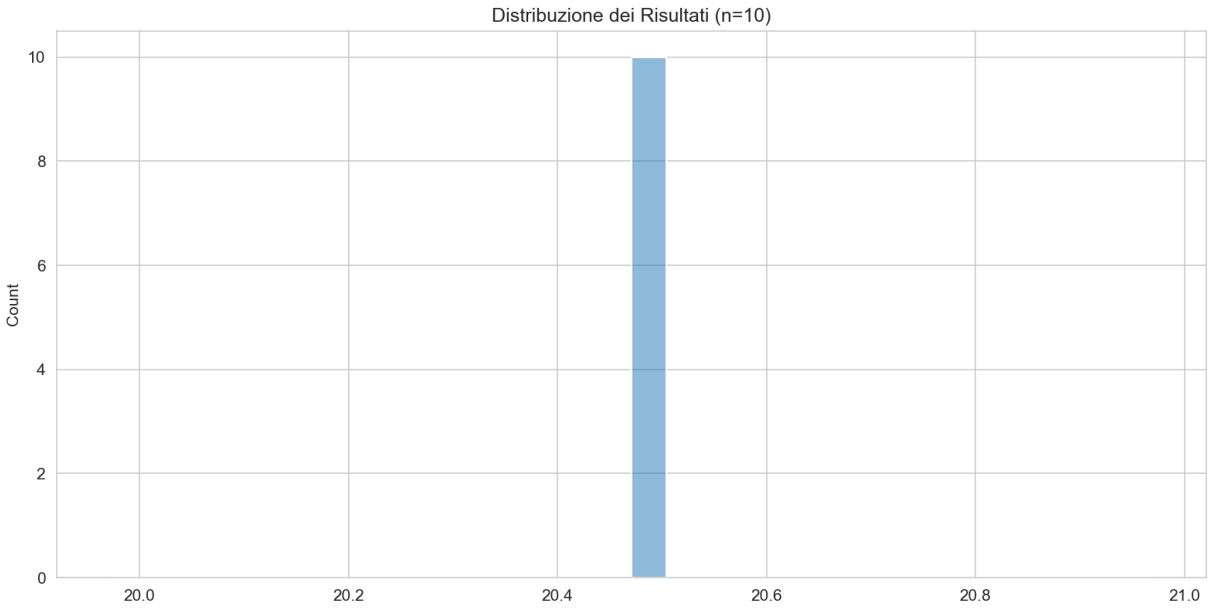
La sensibilità al 67.3%, metrica più scarsa tra tutte quelle valutate, può indicare che il modello sta mancando (falsi negativi) alcuni casi positivi.

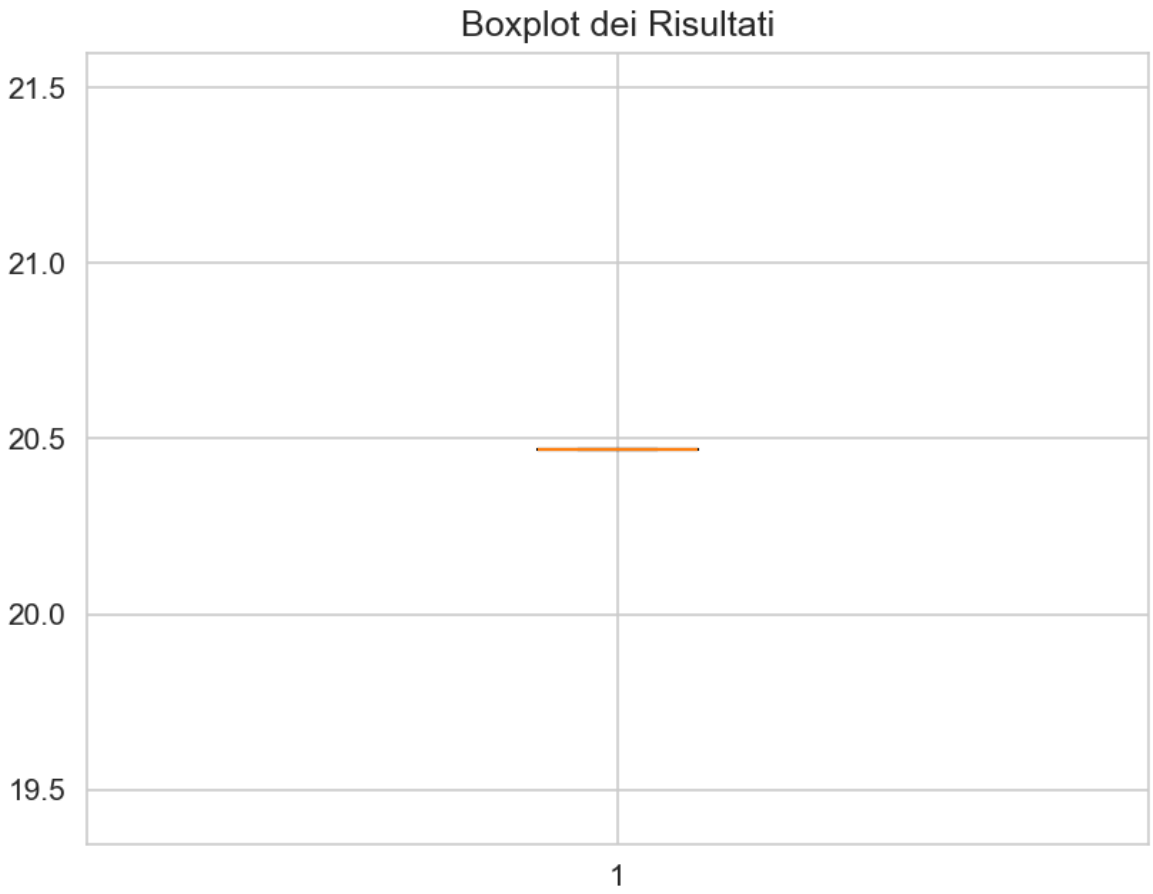
Quindi, possiamo concludere che il modello ha una discreta capacità di identificare correttamente le classi positive (70 su 104) e le classi negative (83 su 96), ma ci sono ancora alcuni errori, specialmente nei falsi negativi (34 casi) che riducono la sensibilità.

Statistica

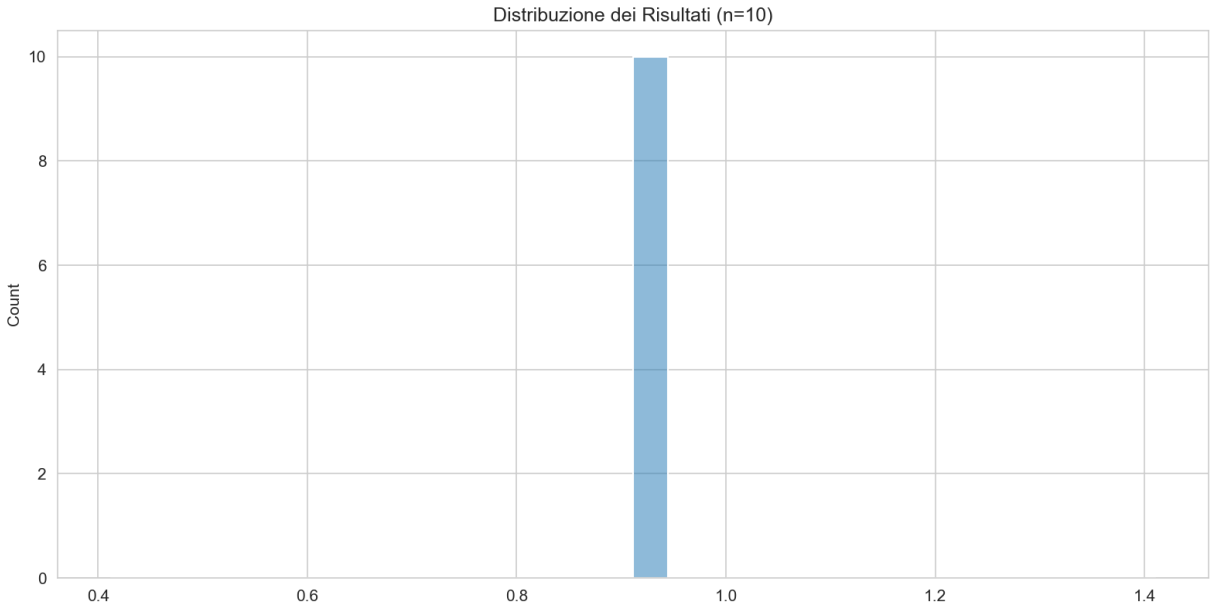
Per quanto riguarda la parte statistica, per avere un risultato più attendibili ho creato uno script che esegue 10 volte i vari modelli (training + test) con lo stesso dataset, salvando i risultati e analizzandoli dal punto di vista della statistica descrittiva (media e diffusione) e della statistica inferenziale (intervallo di confidenza).

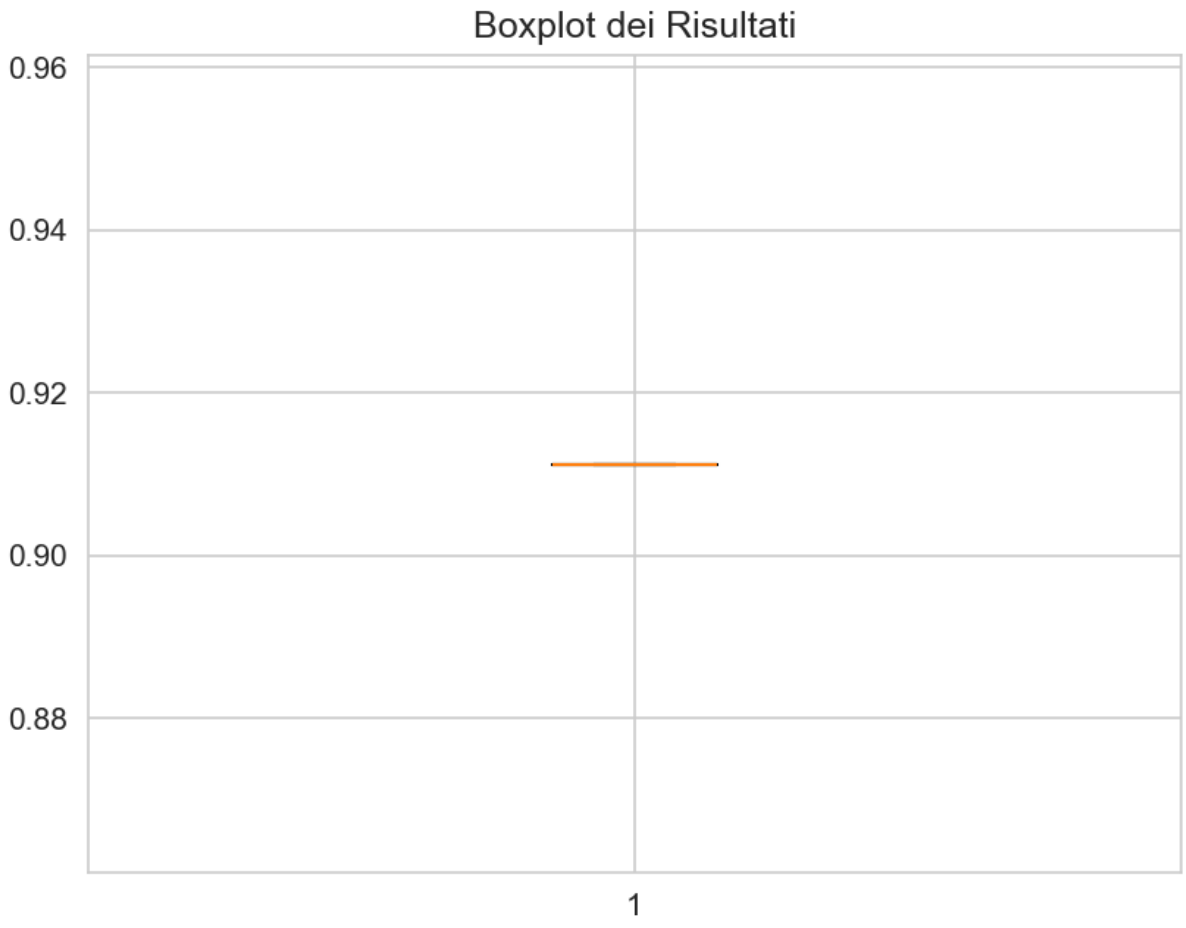
Regressione lineare





MSE: {'means': 20.47086368938935, 'stds': 0.0, 'conf_intervals': (nan, nan)}





R^2: {'means': 0.9112574888913137, 'stds': 0.0, 'conf_intervals': (nan, nan)}

Il valore medio di
MSE è 20.47, indicando che l'errore quadratico medio tra i punteggi previsti e quelli osservati è relativamente basso, inoltre il **R^2** medio di 0.91 suggerisce che il modello di regressione lineare spiega il 91% della variabilità dei dati, indicando una forte capacità predittiva del modello.

Da notare che non ci sono variazioni significative tra le diverse esecuzioni del modello, poiché la **deviazione standard** è 0.0 (a causa di questa mancanza di variabilità, l'intervallo di confidenza non è calcolabile, risultando in valori nan poiché si basa sulla varianza dei dati**)**.

Regressione Logistica

Accuratezza:

- **Media**: 0.7925
- **Deviazione Standard (STD)**: 1.11e-16
- **Intervallo di Confidenza (95%)**: (0.7925, 0.7925)

Misclassification Rate (MR):

- **Media**: 0.2075
- **Deviazione Standard (STD)**: 2.78e-17
- **Intervallo di Confidenza (95%)**: (0.2075, 0.2075)

Precisione:

- **Media**: 0.8040
- **Deviazione Standard (STD)**: 1.11e-16
- **Intervallo di Confidenza (95%)**: (0.8040, 0.8040)

Sensibilità (Recall):

- **Media**: 0.7843
- **Deviazione Standard (STD)**: 0.0
- **Intervallo di Confidenza (95%)**: (NaN, NaN)

Il modello di Logistic Regression mostra una accuratezza media del 79.25%, con una misclassification rate
del 20.75%. Questi valori indicano che il modello è in grado di classificare correttamente circa l'80% delle osservazioni.
Nonostante la deviazione standard sia molto bassa, è diversa da zero per alcune metriche (es. accuratezza e precisione), permettendo il calcolo degli intervalli di

confidenza, che sono molto stretti. (per la sensibilità, non essendoci variazione tra le esecuzioni, non è possibile calcolare un intervallo di confidenza)

SVM (Best Kernel)

Accuratezza:

- **Media**: 0.7650
- **Deviazione Standard (STD)**: 1.11e-16
- **Intervallo di Confidenza (95%)**: (0.7650, 0.7650)

Misclassification Rate (MR):

- **Media**: 0.2350
- **Deviazione Standard (STD)**: 2.78e-17
- **Intervallo di Confidenza (95%)**: (0.2350, 0.2350)

Precisione:

- **Media**: 0.8434
- **Deviazione Standard (STD)**: 0.0
- **Intervallo di Confidenza (95%)**: (NaN, NaN)

Sensibilità (Recall):

- **Media**: 0.6731
- **Deviazione Standard (STD)**: 0.0
- **Intervallo di Confidenza (95%)**: (NaN, NaN)

Il modello SVM ha ottenuto una accuratezza media del 76.5%, leggermente inferiore rispetto alla Logistic Regression, ma ha una precisione superiore (84.34%), ciò lo rende più affidabile nelle predizioni positive corrette (no falsi positivi).
La sensibilità è del 67.31%, il che implica che l'SVM ha più difficoltà a identificare correttamente tutti i veri positivi, con una tendenza a produrre falsi negativi.

Come per la Logistic Regression, gli intervalli di confidenza per accuratezza e MR sono calcolabili grazie a una deviazione standard molto bassa, ma non è possibile calcolare quelli per
sensibilità e precisione, a causa della deviazione standard nulla.

Lorenzo Carnevali