HEART DISEASE PREDICTION DATASET

Lorenzo Maini

INFORMATICA PER IL MANAGEMENT

A.A. 2022-2023

DATASET

Heart Failure Prediction Dataset

11 clinical features for predicting heart disease events.

Facendo Volontariato in Ambulanza mi sono spesso trovato a contatto con pazienti cardiopatici, ed è stata un'occasione unica per fare un'indagine approfondita su questo tema e capire meglio i diversi sintomi e le loro cause.

Il dataset scelto riporta 11 features di pazienti ricoverati per scompenso cardiaco, per un totale di 918 osservazioni fatte in diversi ospedali del mondo. Tra queste features era presente anche la presenza o meno di malattie cardiache pregresse, che poi ho deciso di usare come target per il mio modello.

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

Final dataset: 918 observations

PRE-PROCESSING

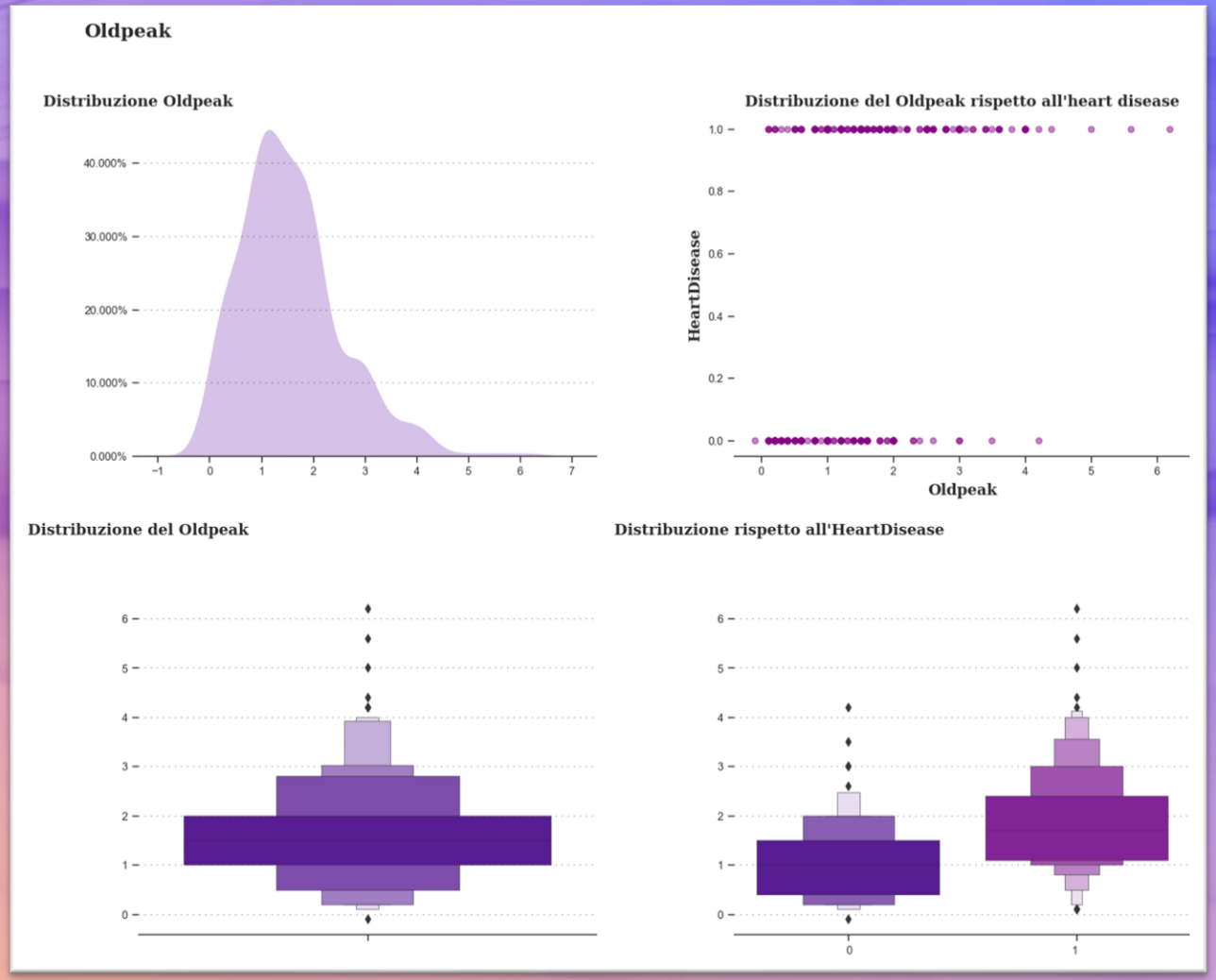
In questa fase sono stati rimossi alcuni Nan, e sono stati tolti dei valori che, pur essendo segnalati al 100% validi su Kaggle, non erano attendibili. In questo modo il Dataset è passato da 918 a 429 righe.
Inoltre, per facilitare e rendere più completa la parte di Esplorazione successiva, sono state sostituite tre features stringhe in numeri interi, ovvero Sesso, Pendenza del Segmento ST, e l'Angina Causata dell'Esercizio.


```
In [8]:
   ...:
   ...: print("Numero di righe nel dataset dopo aver eliminato delle righe con valori
uguali 0:", len(df))
   ...: df.info()
Numero di righe nel dataset dopo aver eliminato delle righe con valori uguali 0: 429
<class 'pandas.core.frame.DataFrame'>
Int64Index: 429 entries, 1 to 915
```

```
51
52  #Cambio le stringhe in valori numerici
53  df['Sex'] = df['Sex'].replace({'M': 1, 'F': 0})
54  df['ExerciseAngina'] = df['ExerciseAngina'].replace({'Y': 1, 'N': 0})
55  df['ST_Slope'] = df['ST_Slope'].replace({'Up': 0, 'Flat': 1,'Down':2})
56
```

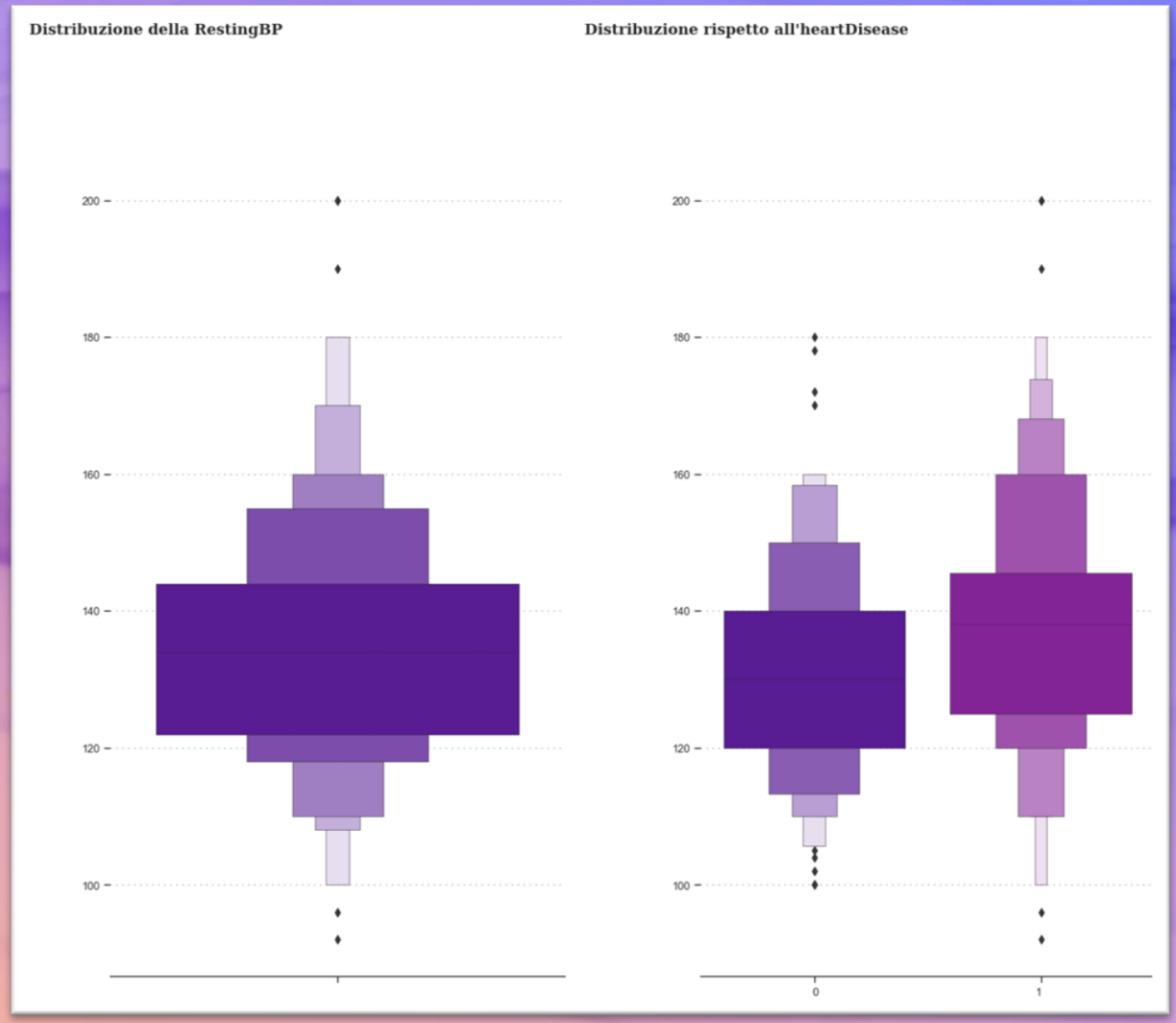
EDA

- **Distribuzione dell’Oldpeak**
- **Distribuzione dell’Oldpeak rispetto a HeartDisease**



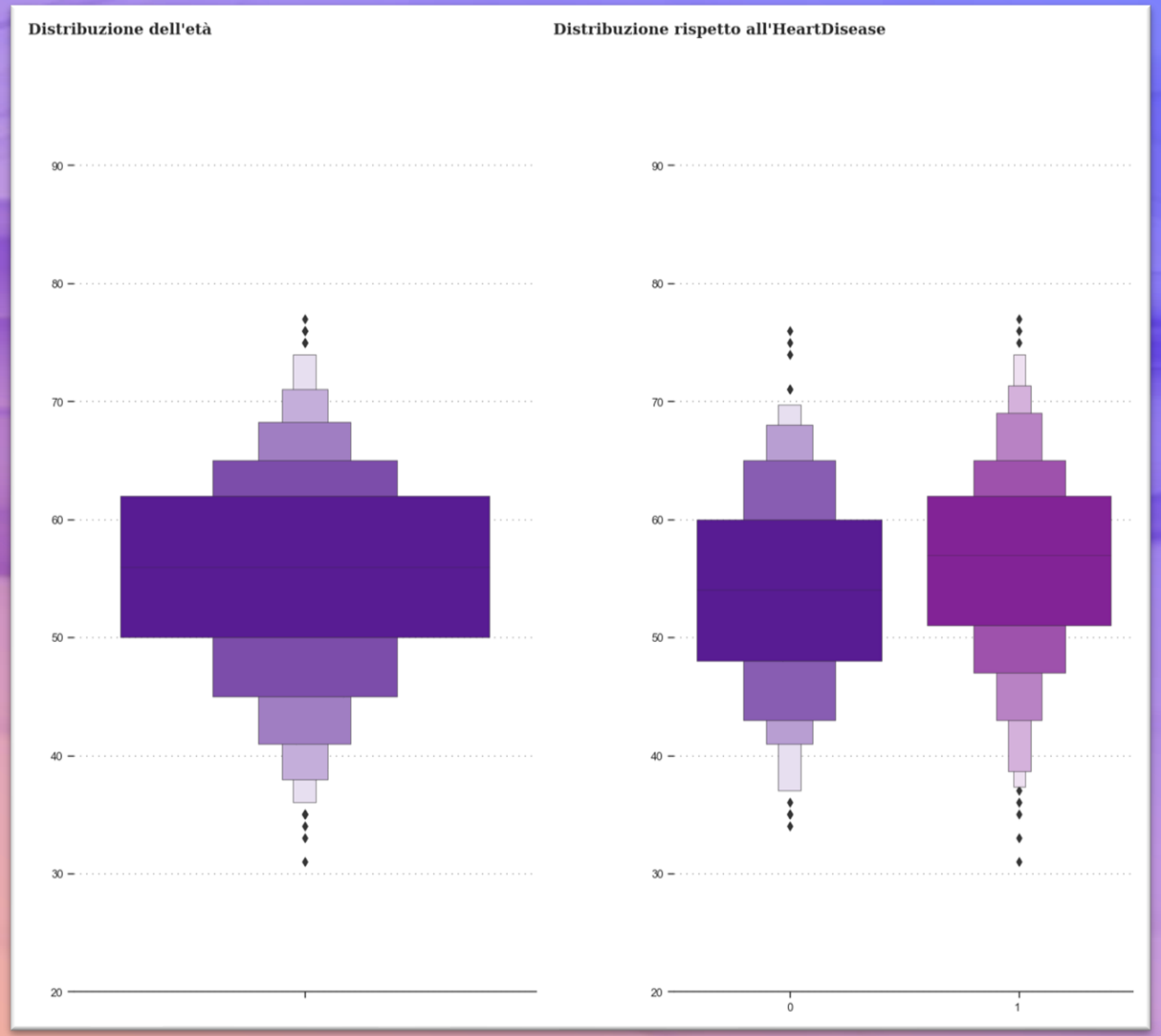
EDA

- **Distribuzione della Resting Blood Pressure**
- **Distribuzione della Resting Blood Pressure rispetto ad Heart Disease**



EDA

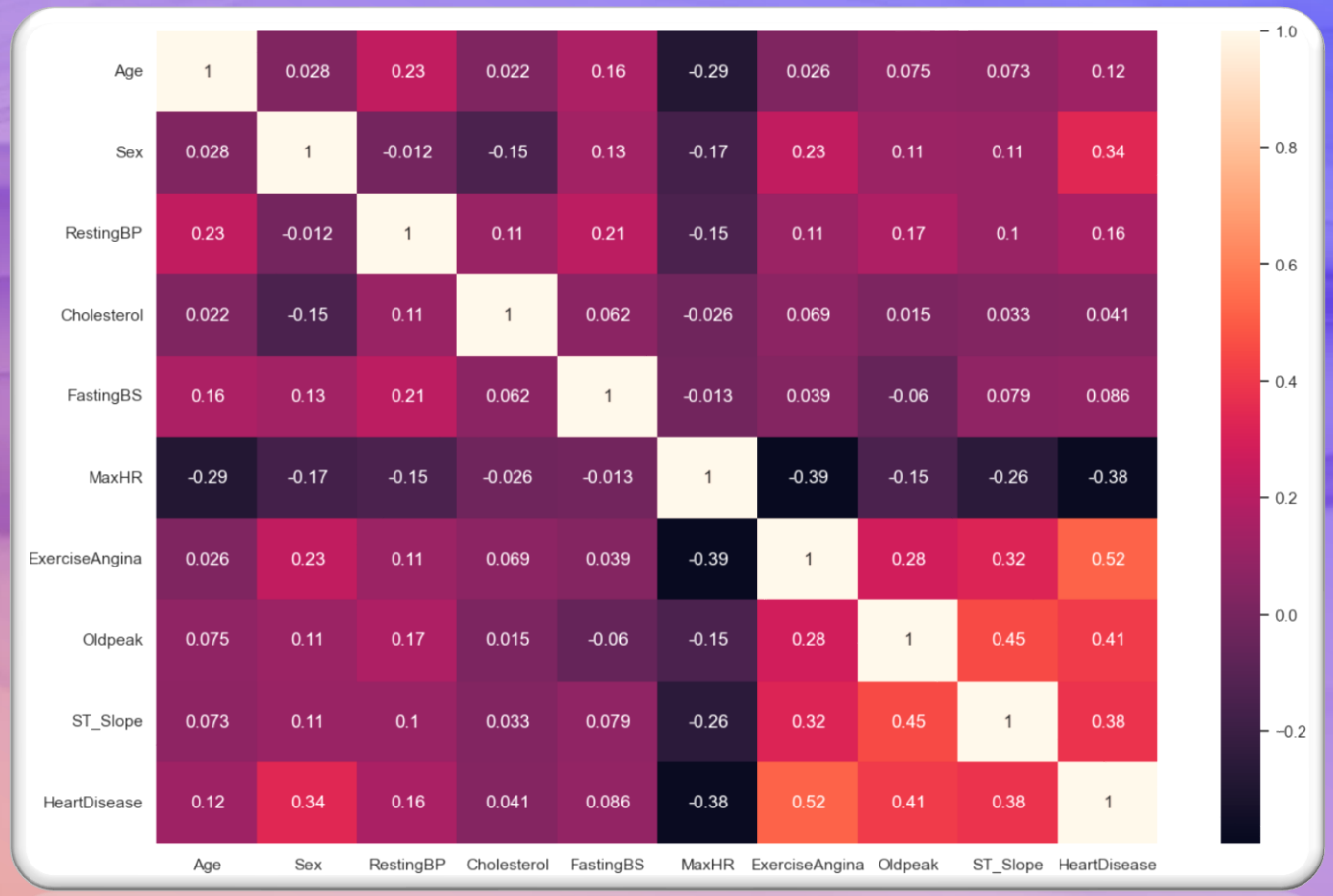
- **Distribuzione dell’età**
- **Distribuzione dell’età rispetto a Heart Disease**



EDA

Tramite la matrice di correlazione possiamo notare alcuni aspetti interessanti:

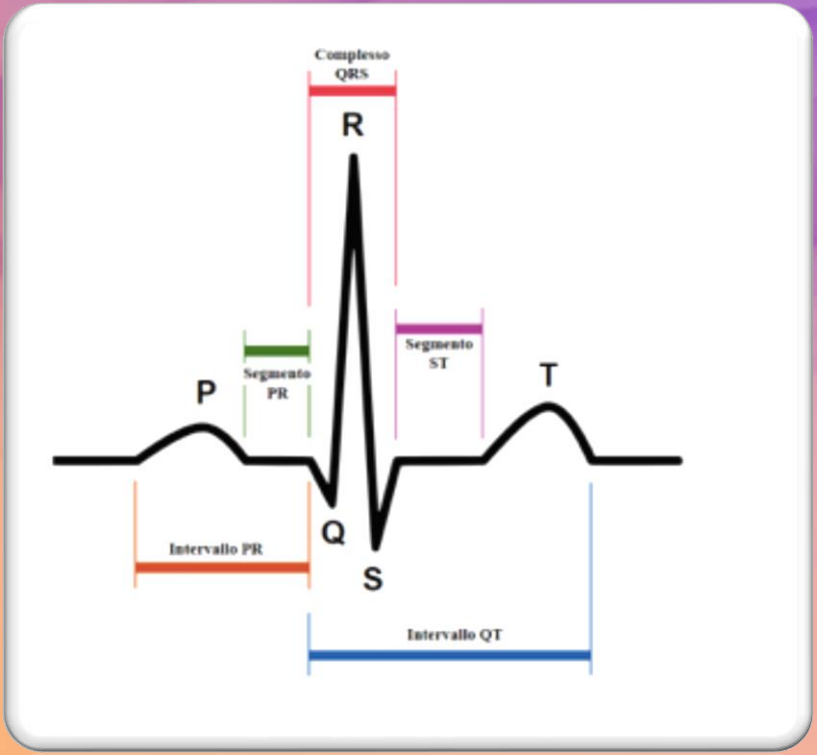
- **Correlazione Positiva con ExerciseAngina, ovvero se il dolore è cominciato durante esercizio fisico**
- **Correlazione Positiva con OldPeak ed ST_Slope**
- **Correlazione Positiva con il Sesso (variabile convertita da string a intero -> M=1, F=0)**
- **Correlazione Negativa con MaxHR, cioè la frequenza cardiaca massima del paziente**

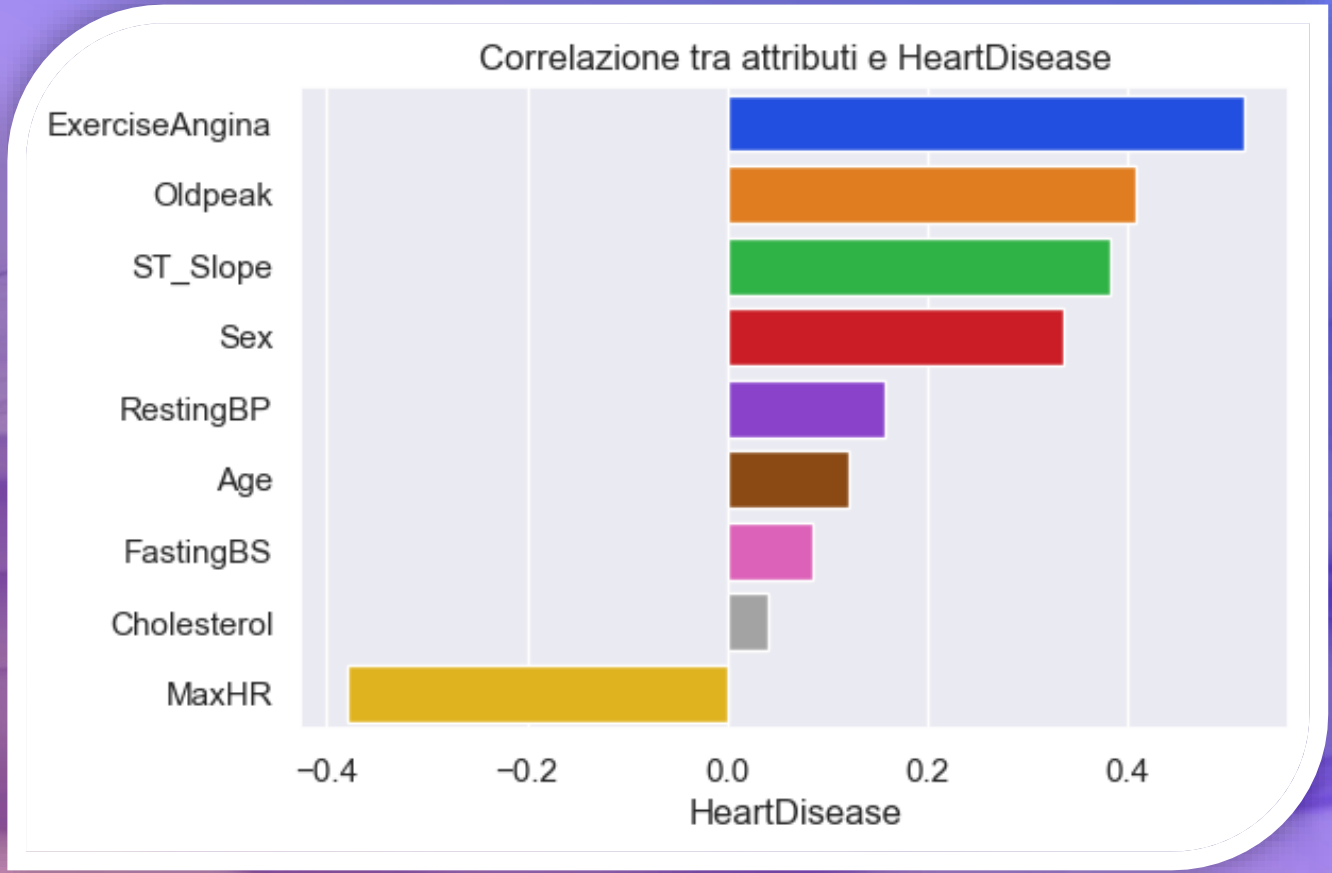


EDA

OLDPEAK ed ST SLOPE

L'oldpeak nel database indica il valore numerico della pendenza del tratto ST dell'ECG, mentre ST_Slope indica la stessa cosa ma con una stringa (Flat=0, Up=1, Down=2), che è stata convertita ad intero.





L'elettrocardiogramma (E.C.G) è la principale e più diffusa metodica di indagine cardiologica, fornisce la rappresentazione grafica dell'attività elettrica del cuore durante il suo funzionamento. L'E.C.G. permette di identificare la presenza di disturbi del ritmo cardiaco.

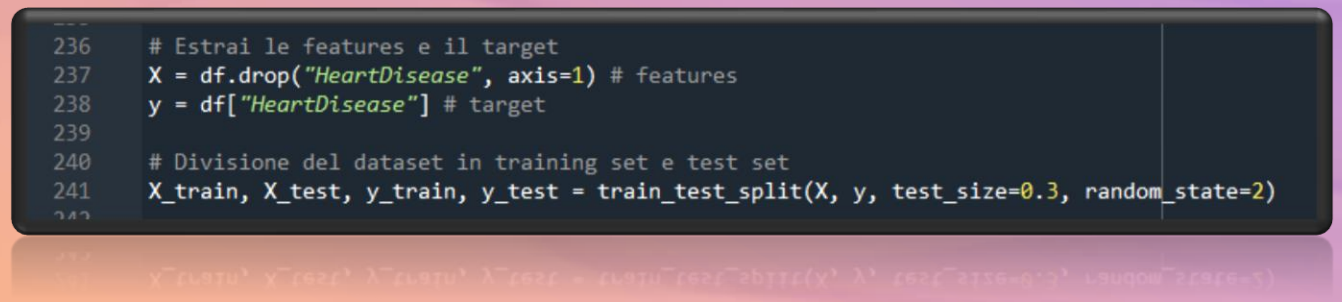
Il tratto ST è quello in cui avviene l'intervallo tra la depolarizzazione e la ripolarizzazione ventricolare, cioè quando si stabilizzano le condizioni elettriche di base. Se il tratto non è piatto, spesso si hanno problemi cardiaci nel paziente.

SPLITTING

Il Dataset successivamente è stato suddiviso in train set e test set con una proporzione 70-30. E' stata provata anche la divisione 80-20 ma ha portato ad una precisione inferiore.
Delle 9 Colonne è stata scelta «Heart Disease» come target e le altre come features.
Poi, sono stati applicati i due algoritmi di machine learning per addestrare il modello.

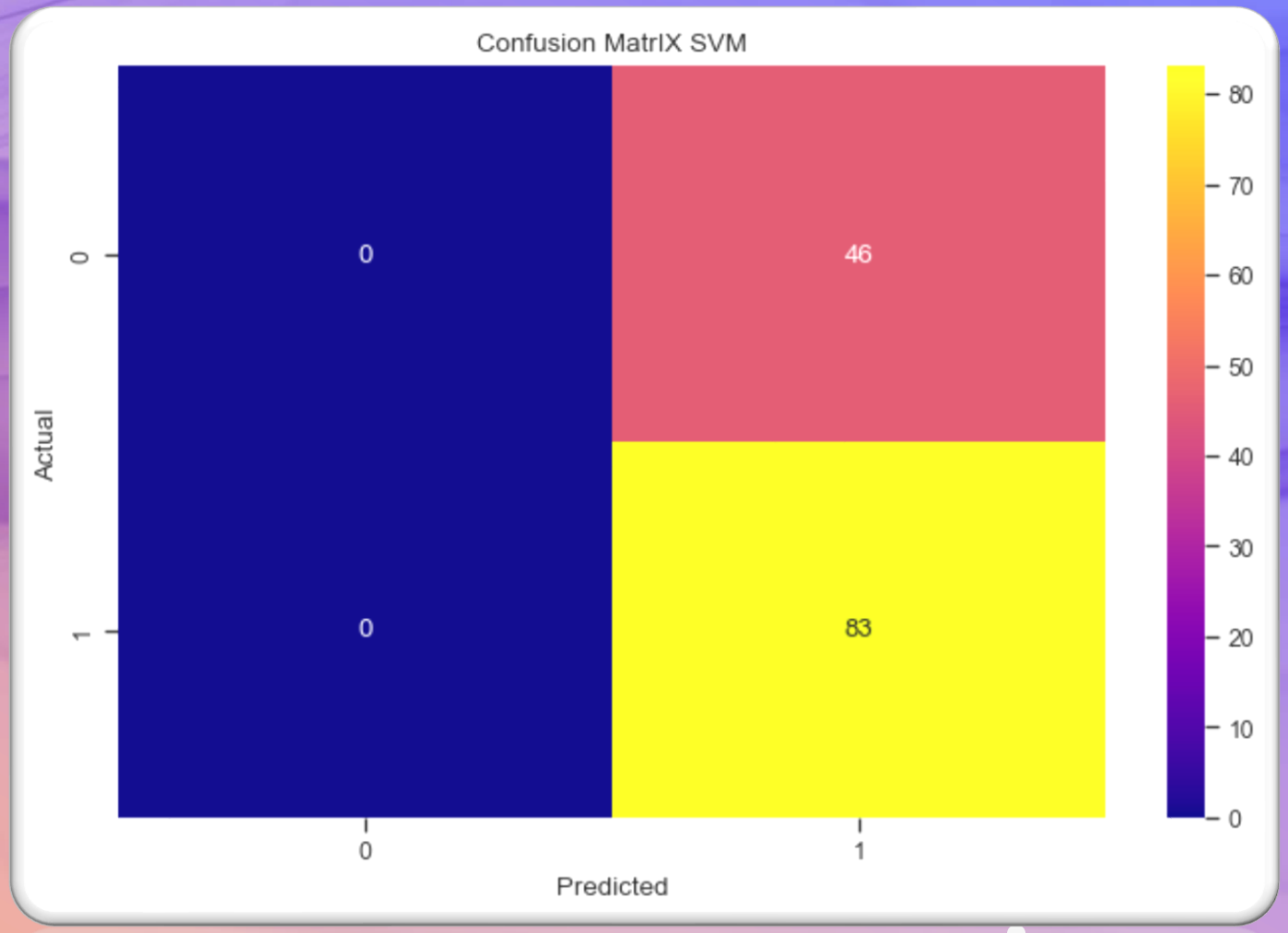
- **Support Vector Machine**

- **Regressione Logistica**



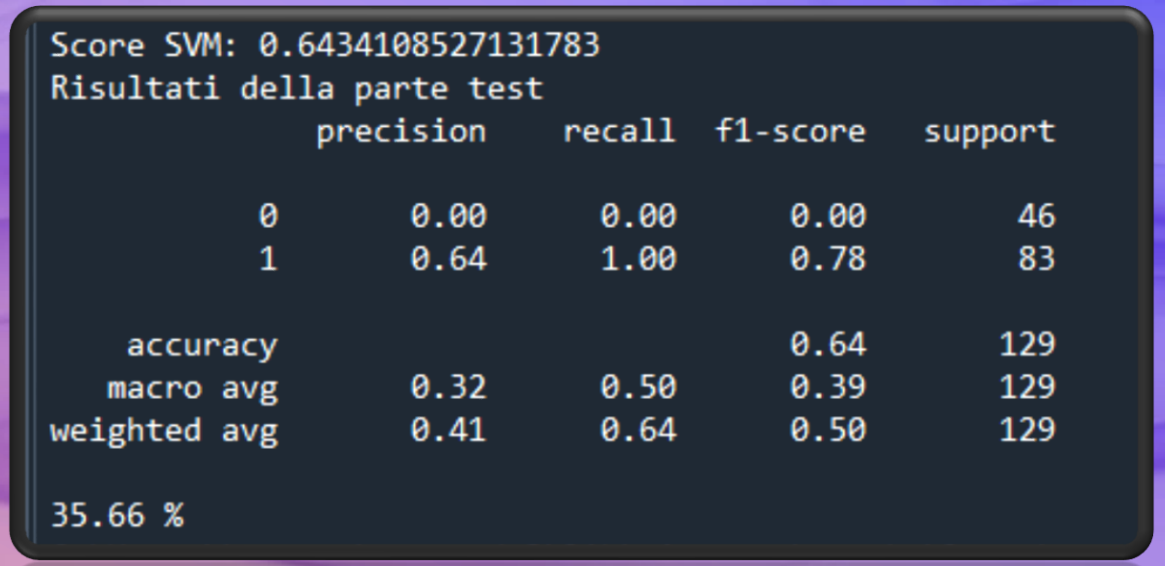
SVM

- **Il modello ha previsto correttamente la maggior parte dei dati. Tuttavia, notiamo che non ha previsto alcun paziente sano, quando ne erano presenti 46.**



RISULTATI SVM

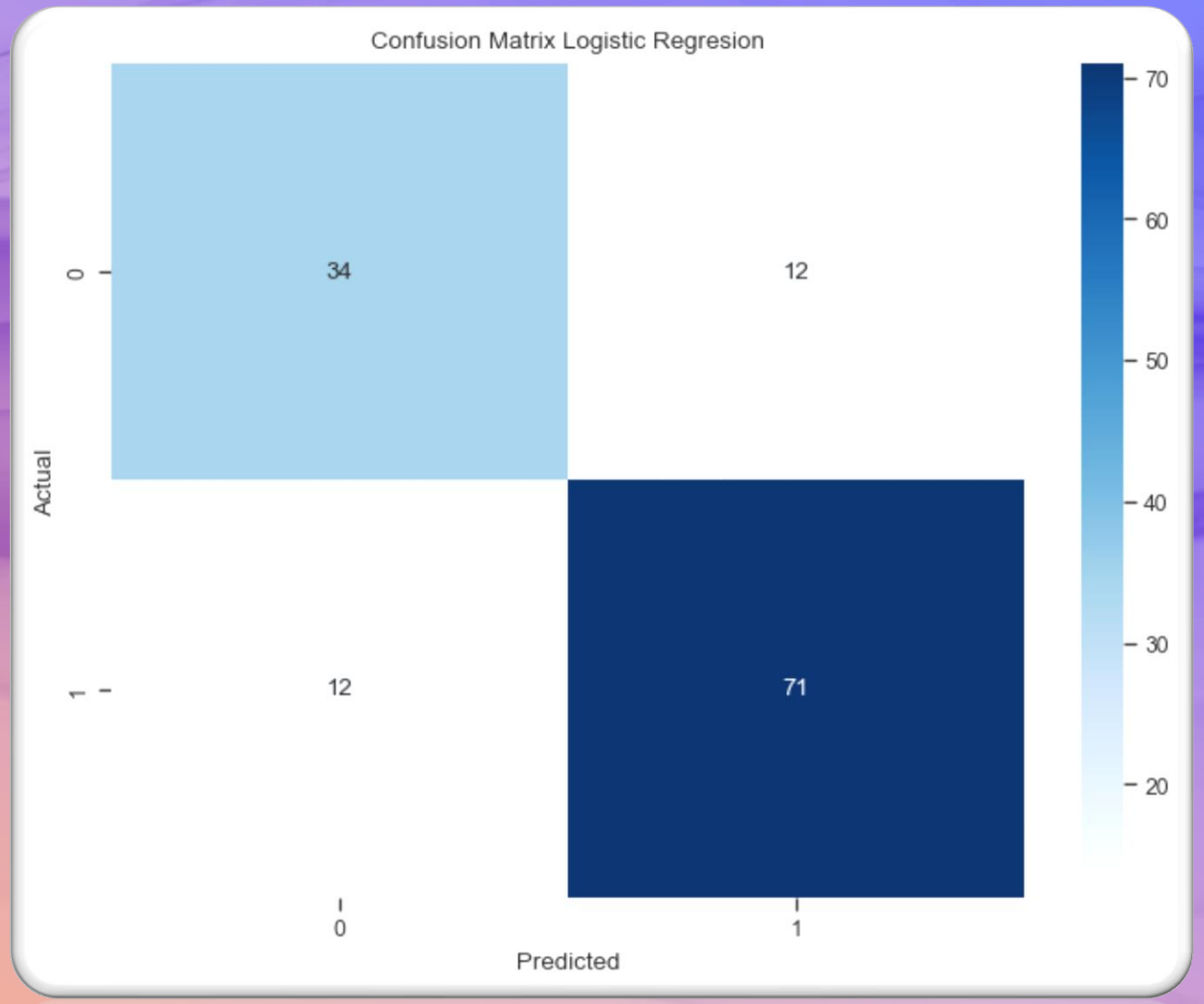
- Tabella dei Risultati con SVM
- “Precision” : istanze classificate correttamente come positive rispetto a tutte le istanze classificate come positive
- “Recall” : istanze positive classificate correttamente rispetto a tutte le istanze effettivamente positive (1.00 indica che tutte le istanze positive sono state identificate correttamente dal modello)
- “F1-score ” rappresenta una media ponderata della “ precisione ” e del “ recall” (1.00 = bilanciamento perfetto tra precisione e recall)
- “Support” indica il numero di campioni di ogni classe nel set di dati di test
- Miscalculation Rate: 35.66%



REGRESSIONE LOGISTICA

- **Il modello, utilizzando la regressione logistica, ha previsto correttamente un numero maggiore di dati rispetto al precedente caso. Ha avuto infatti un'accuratezza dell'81% ed un miscalculation rate del 19%, con previsione anche più eterogenee rispetto a prima.**

```
  n_iter_i = _check_optimize_result(
Accuracy: 0.813953488372093
18.6 %
```

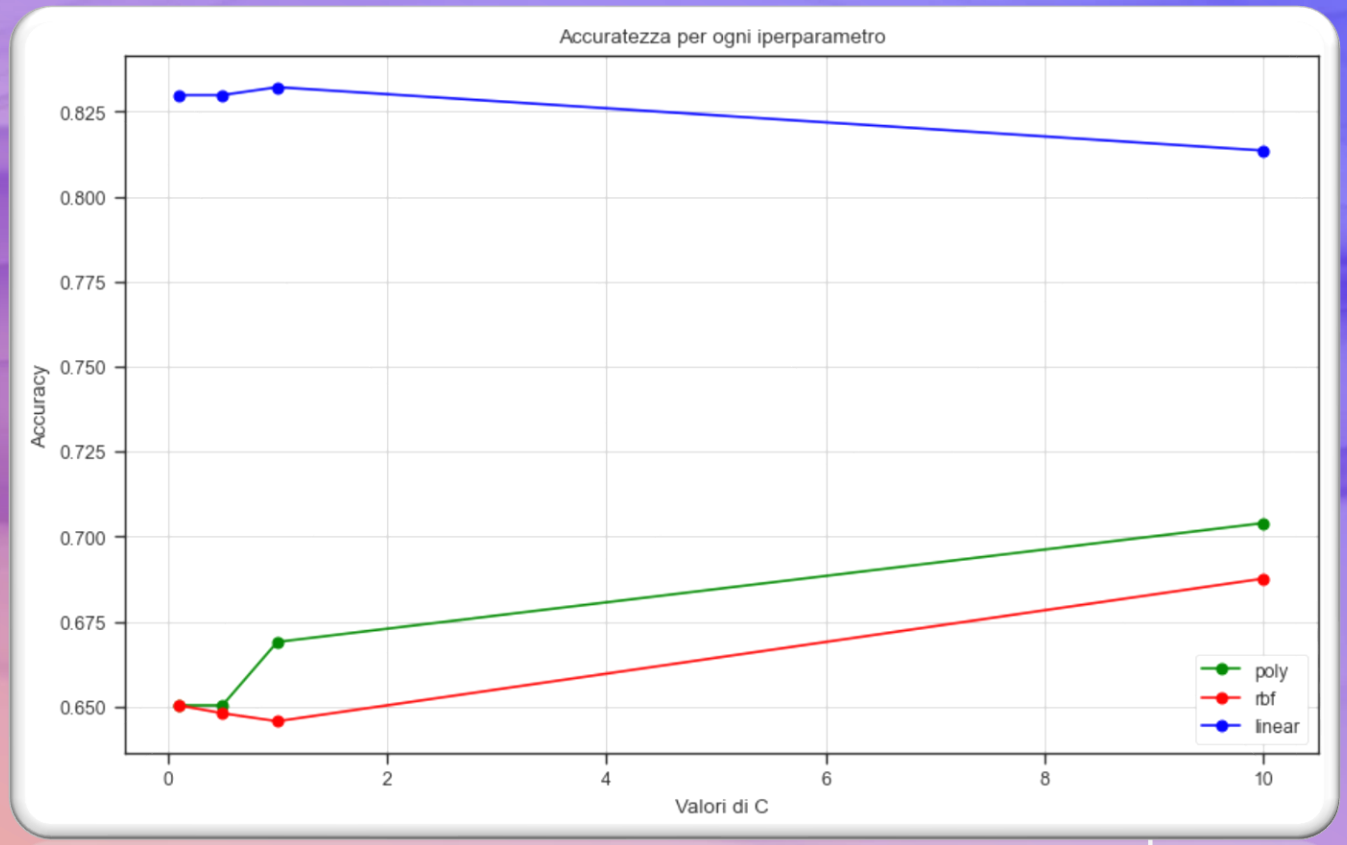


HYPERPARAMETER TUNING (SVM)

- **Le performance del modello dipendono dalla scelta degli iperparametri.**
- **Con il processo di tuning dei parametri, sono stati trovati gli iperparametri ideali per il modello.**
- **Si tratta di {'C': 1, 'kernel': 'linear'}.**

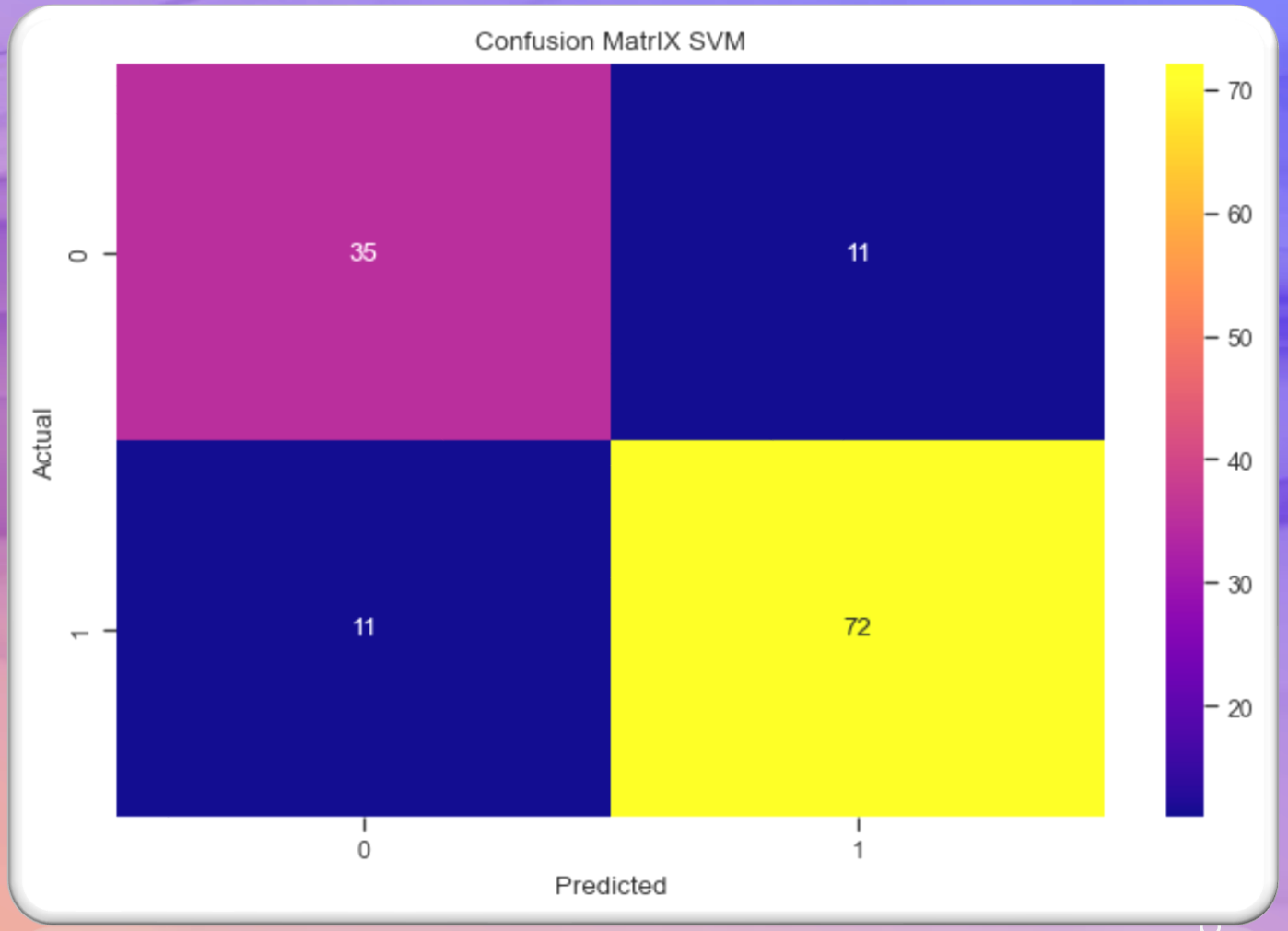
```
   ...: plt.show()
Miglior set di iperparametri trovato:
{'C': 1, 'kernel': 'linear'}
Miglior accuratezza: 0.8321678321678323

In [10]:
```



SVM CON IPERPARAMETRI OTTIMIZZATI

- **Questa la matrice di confusione utilizzando SVM con iperparametri ottimizzati, visibilmente migliore della precedente.**



RISULTATI SVM HP OTTIMIZZATI

- **Con la stessa tabella di prima individuiamo il netto miglioramento della precisione, ovvero 83%, con un miscalculation rate del 17%.**

```
   ...: print(MK, '%')
Miglior set di iperparametri trovato:
{'C': 1, 'kernel': 'linear'}
Miglior accuratezza: 0.8321678321678323
Score SVM: 0.8294573643410853
Risultati della parte test
              precision    recall  f1-score   support

           0       0.76      0.76      0.76        46
           1       0.87      0.87      0.87        83

    accuracy                           0.83       129
   macro avg       0.81      0.81      0.81       129
weighted avg       0.83      0.83      0.83       129

17.05 %
```

STUDIO STATISTICO SULLE PERFORMANCE

BOXPLOT OTTIMIZZATI

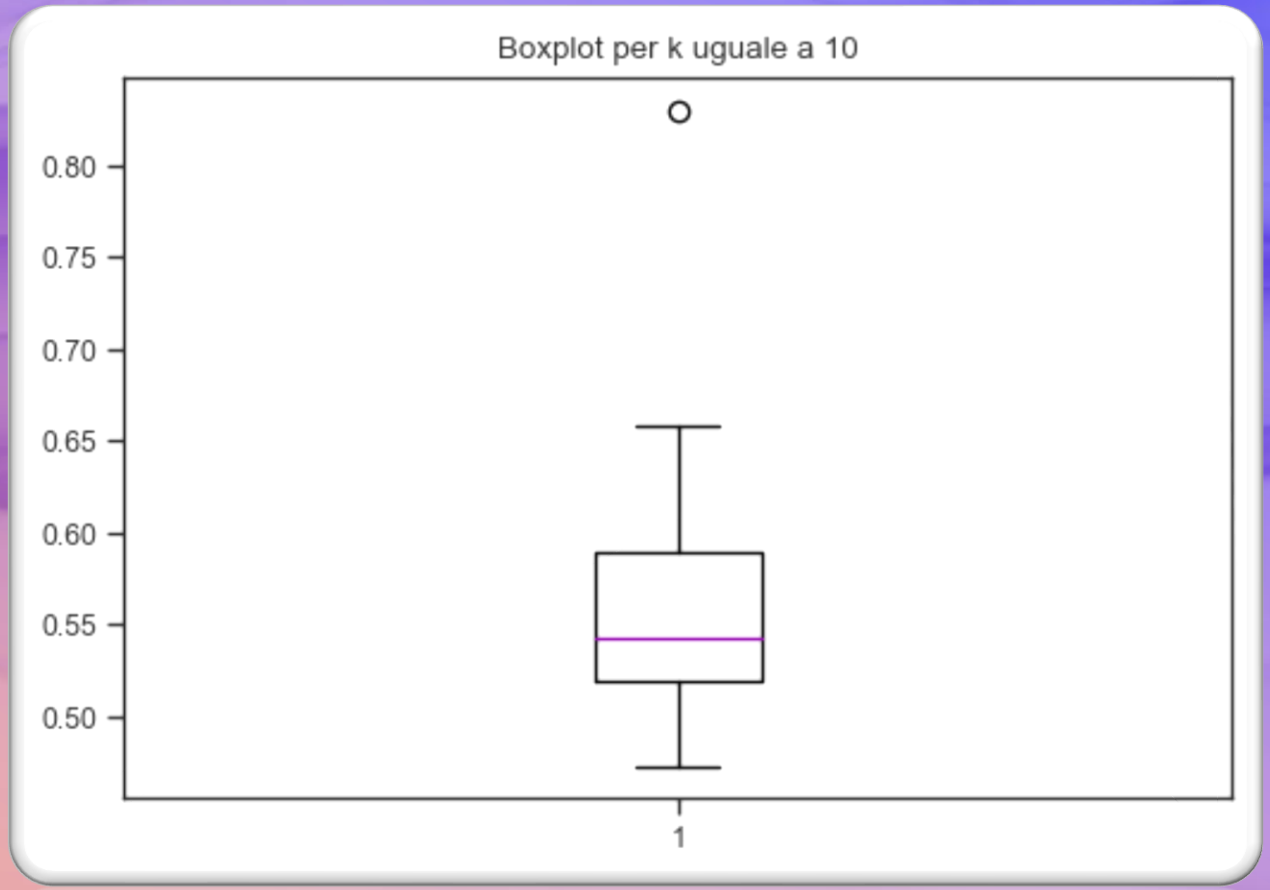
- Successivamente si è eseguito l'addestramento e il testing del modello utilizzando
gli iperparametri ottimizzati e un valore di k maggiore uguale a 10, per valutare in modo più accurato il modello.

Per k=10, otteniamo i seguenti risultati:

Mean Accuracy: 0.5751937984496125

Standard Deviation of Accuracy: 0.09828772626725868

Confidence Interval (95%): (0.5010797037879441, 0.6493078931112809)



STUDIO STATISTICO SULLE PERFORMANCE

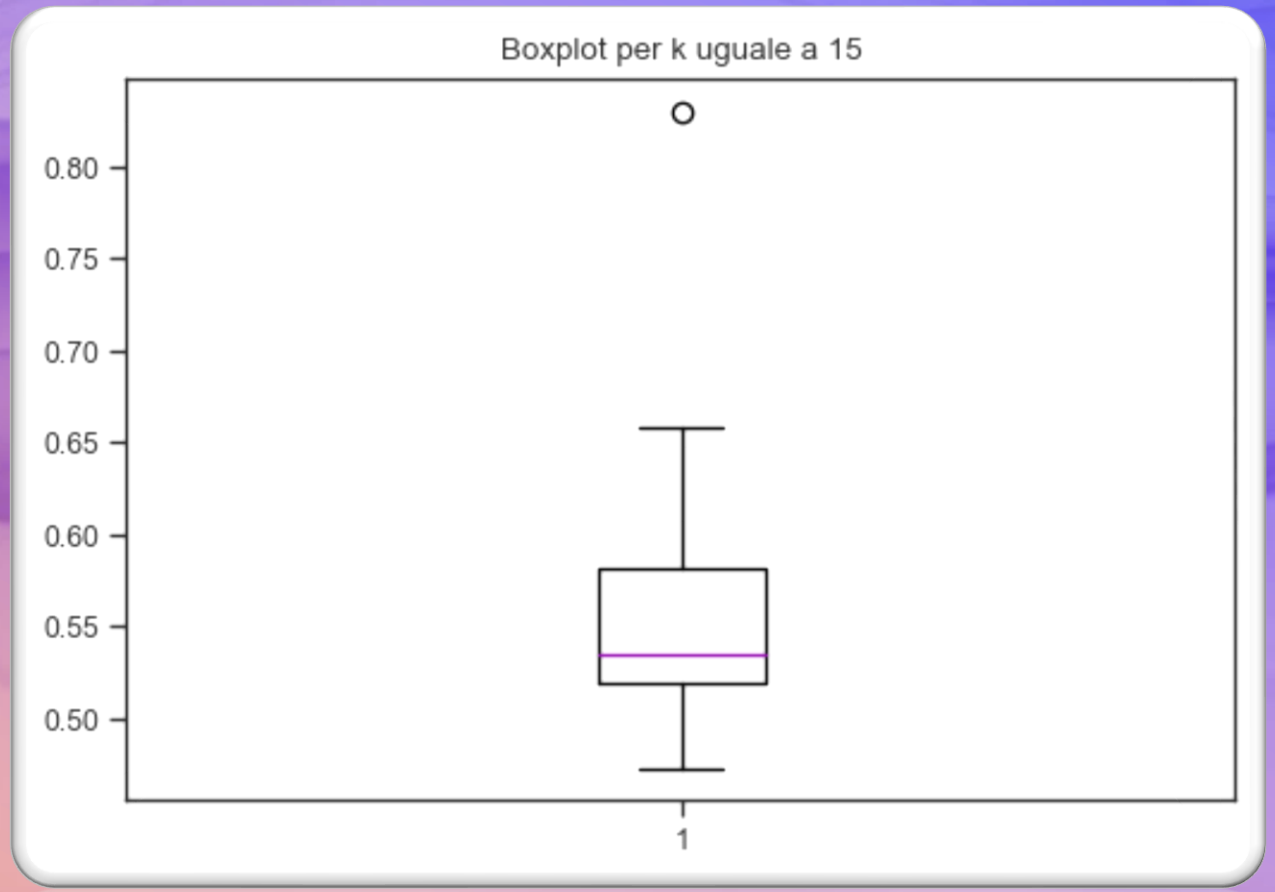
BOXPLOT OTTIMIZZATI

Per k=15, otteniamo i seguenti risultati:

Mean Accuracy: 0.5643410852713179

Standard Deviation of Accuracy: 0.08469147462789248

Confidence Interval (95%): (0.5157943775999957, 0.6128877929426401)



STUDIO STATISTICO SULLE PERFORMANCE

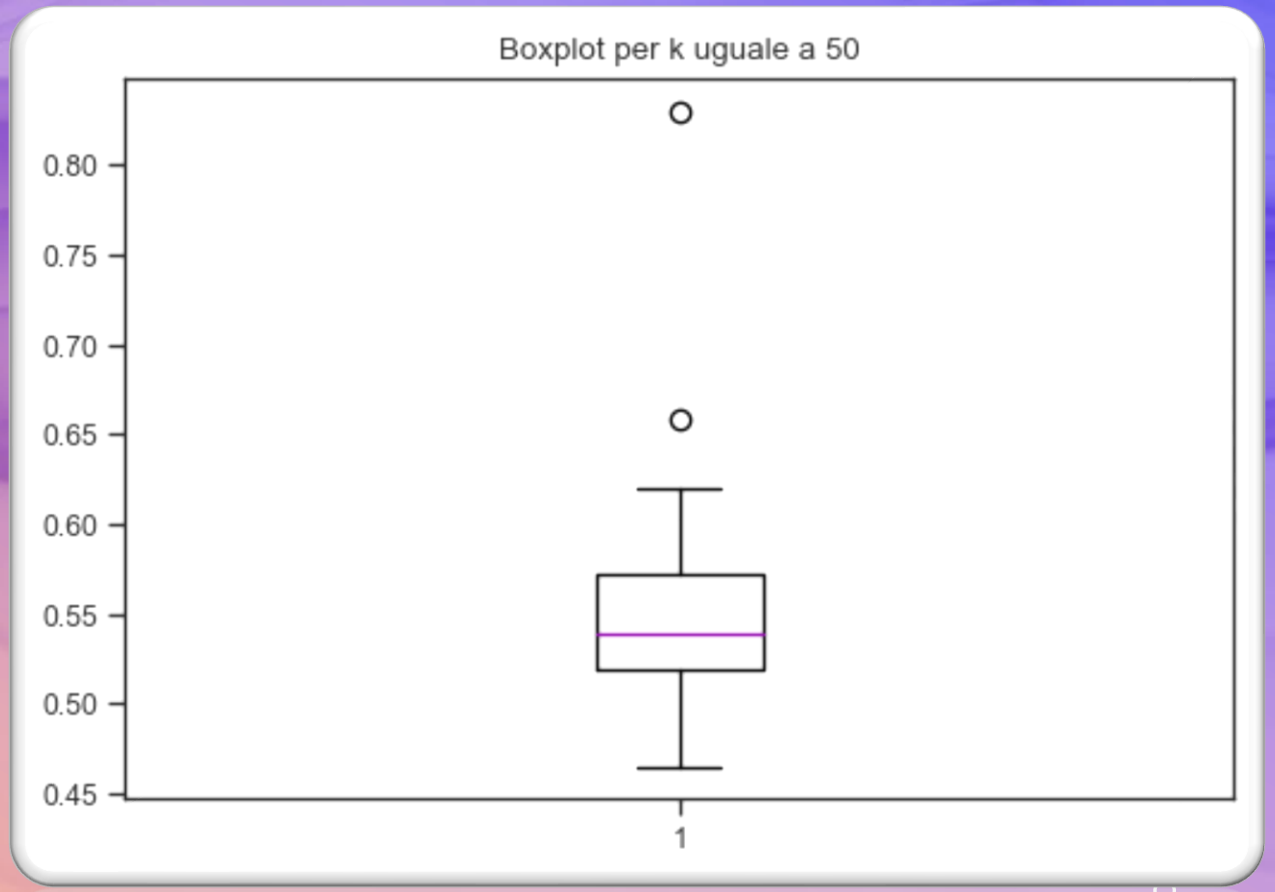
BOXPLOT OTTIMIZZATI

Per k=50, otteniamo i seguenti risultati:

Mean Accuracy: 0.5483720930232558

Standard Deviation of Accuracy: 0.057673793444973885

Confidence Interval (95%): (0.5318149748823894, 0.5649292111641223)



STUDIO STATISTICO SULLE PERFORMANCE

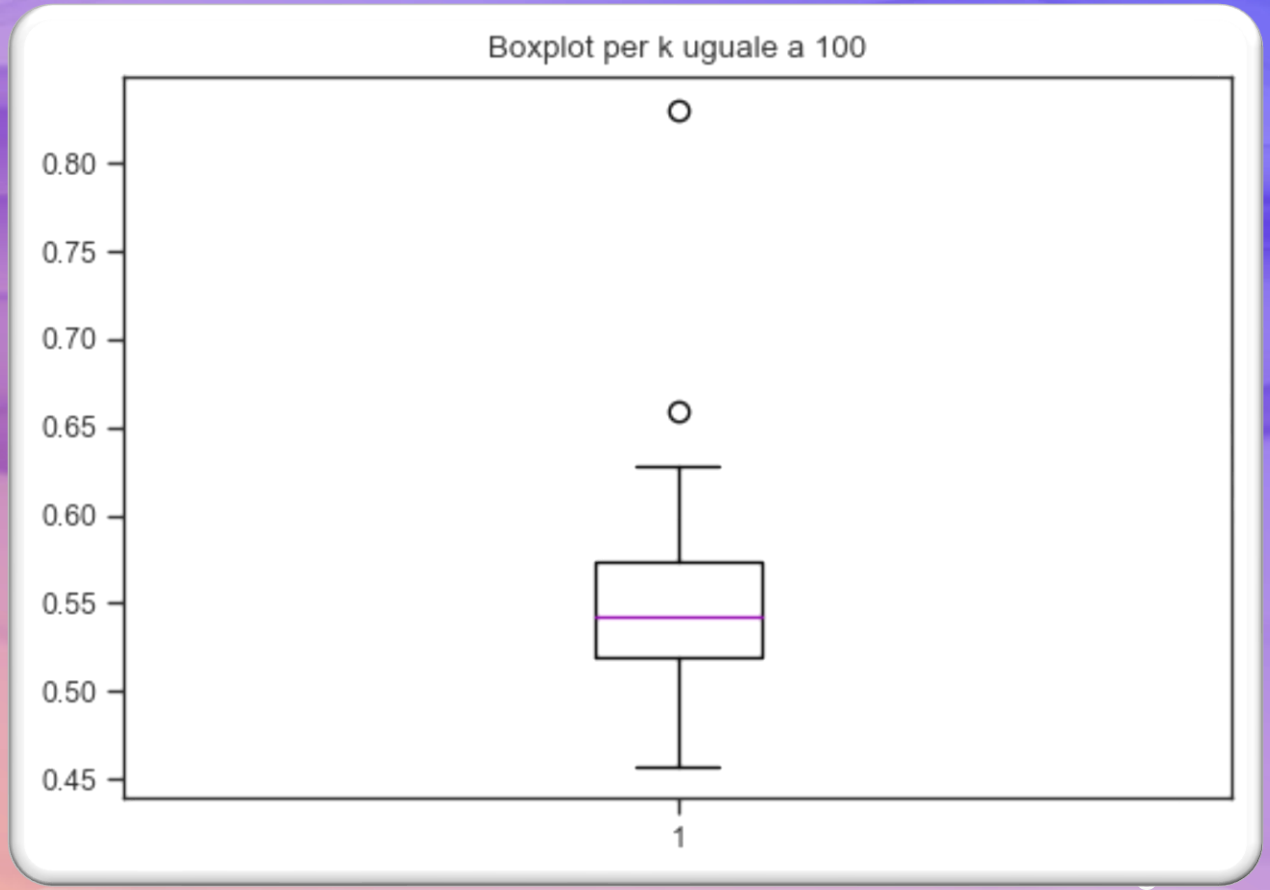
BOXPLOT OTTIMIZZATI

Per k=100, otteniamo i seguenti risultati:

Mean Accuracy: 0.5470542635658915

Standard Deviation of Accuracy: 0.0502652659330212 7

Confidence Interval (95%): (0.5370302985371044, 0.5570782285946786)



FINE