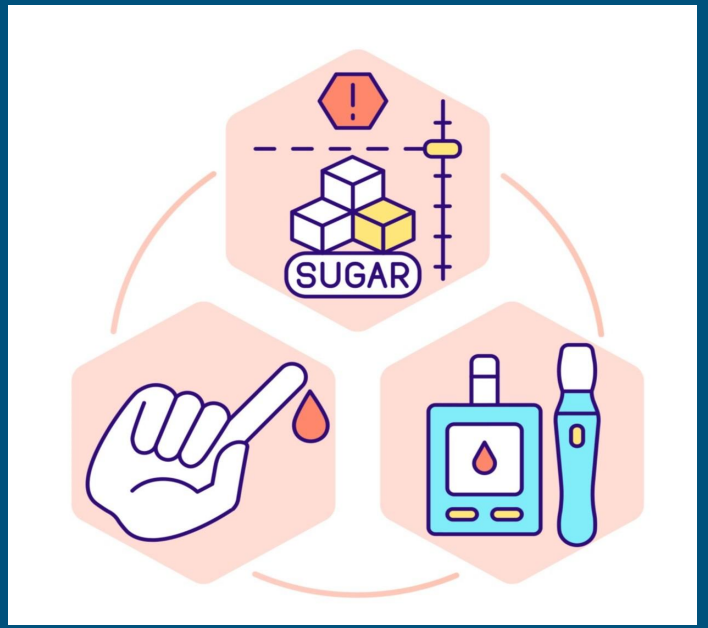
Diabetes Dataset

Informazioni sul Dataset

L'obiettivo di questo dataset è la diagnosi predittiva del diabete basata su misurazioni diagnostiche. Sono state applicate limitazioni alla selezione delle istanze, i dati sono stati raccolti da donne di origine indiana di almeno 21 anni di età.

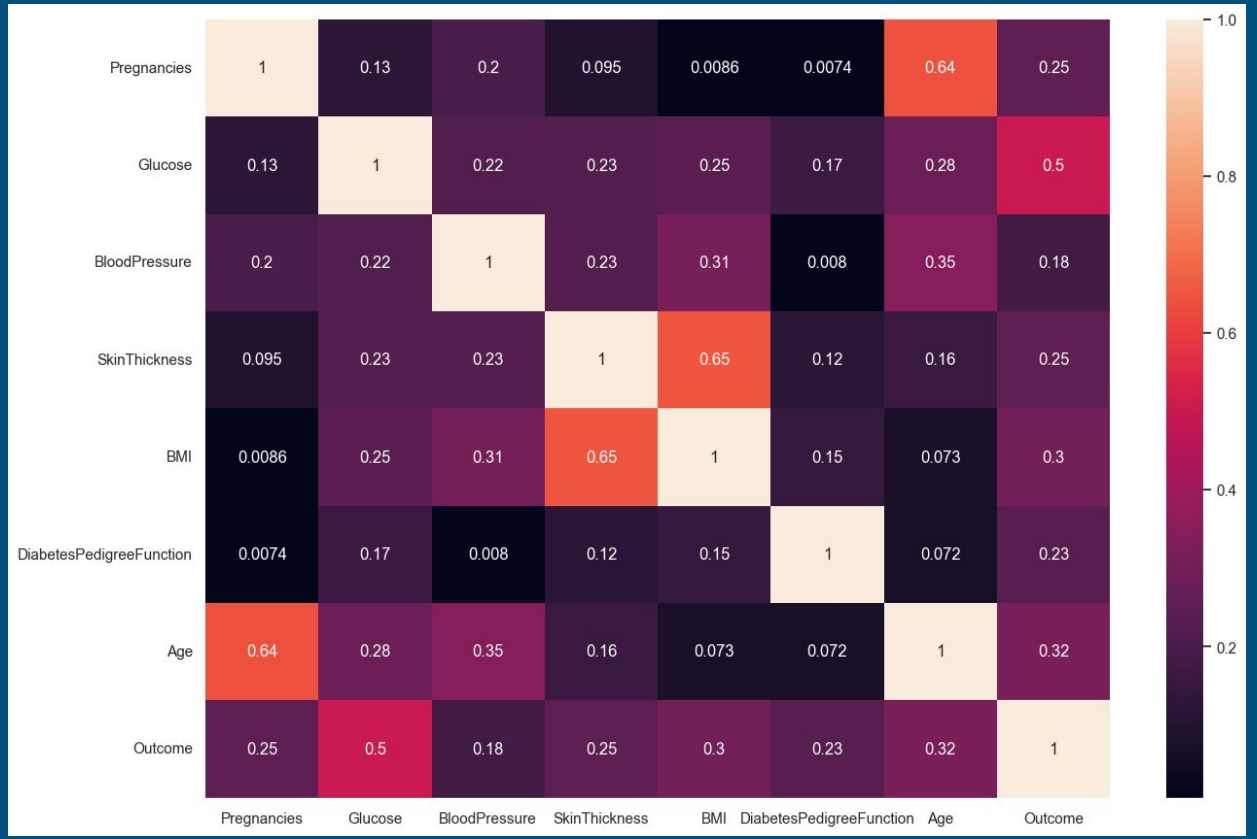
La variabile target chiamata è denominata "Outcome",ed indica se la persona in questione è diabetica oppure no.

Sono stati rilevati dei dati corrotti.



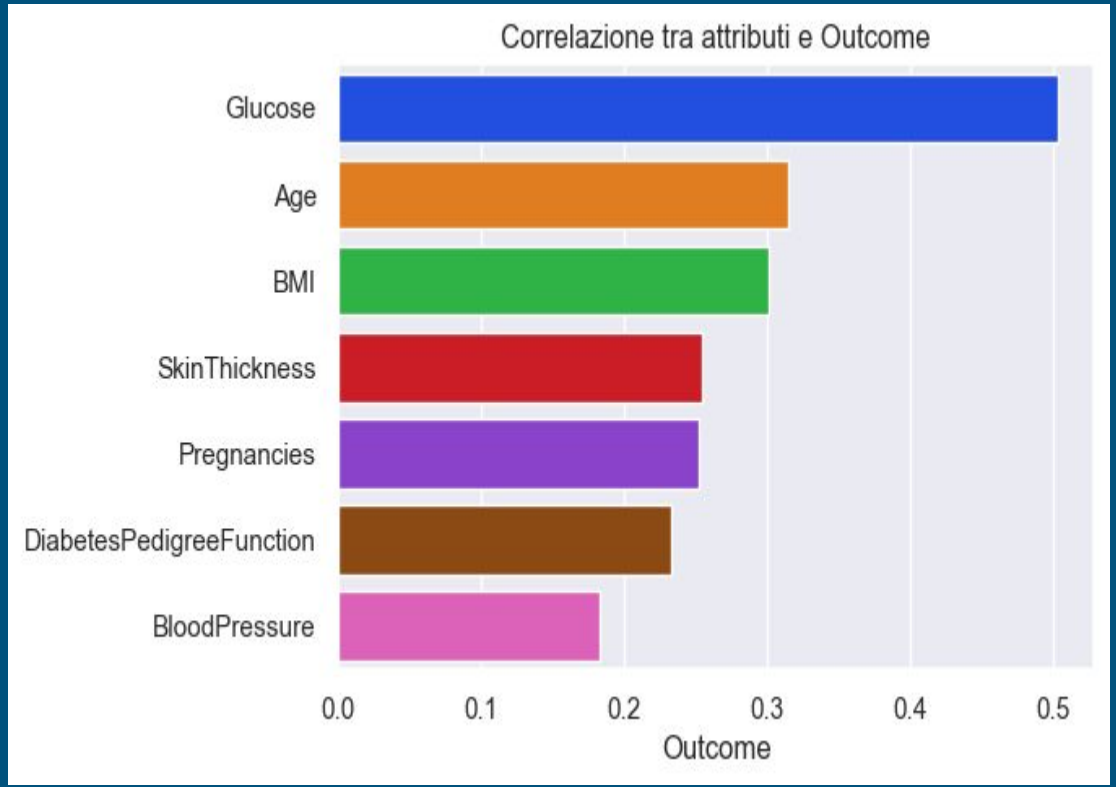
Exploratory Data Analysis

La matrice di correlazione del dataset evidenzia una correlazione significativa tra l'outcome e il livello di glucosio della paziente (0.5), indicando un'associazione rilevante tra queste due variabili.



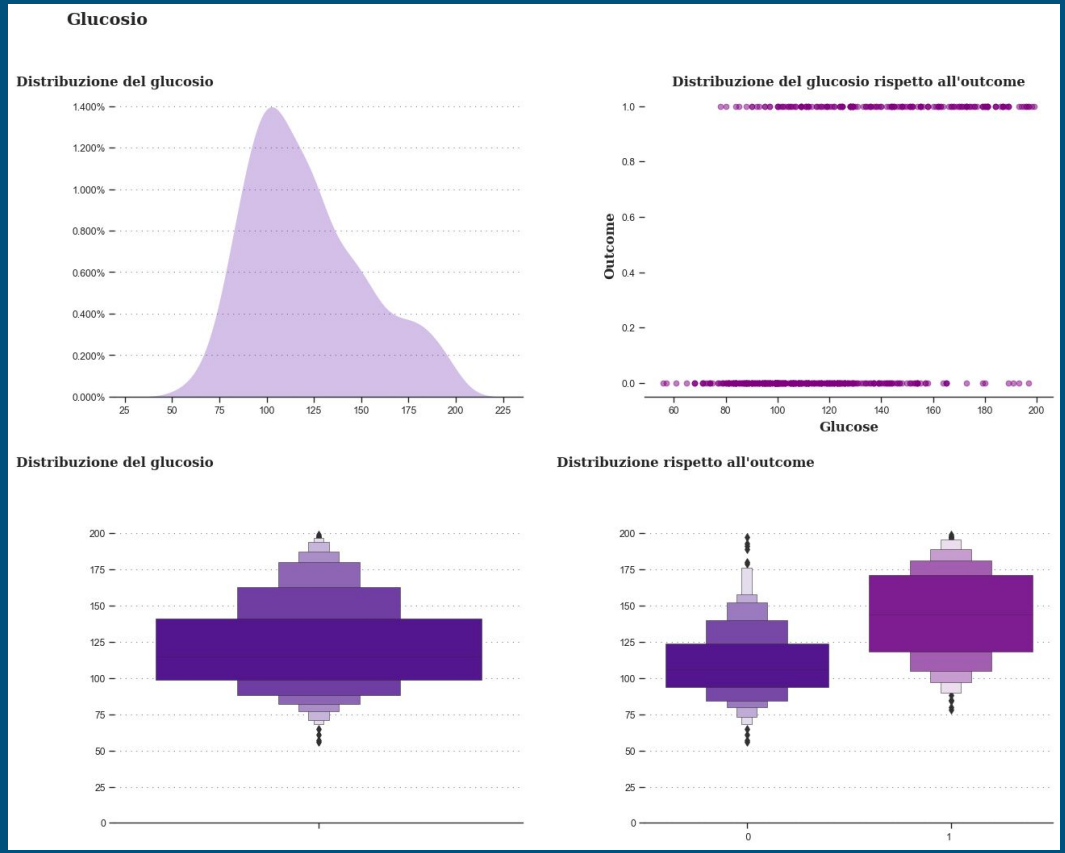
Exploratory Data Analysis

In questa immagine sono illustrate le correlazioni dell'outcome con le diverse variabili del dataset, fornendo una panoramica delle relazioni esistenti tra l'outcome e ciascuna variabile considerata.



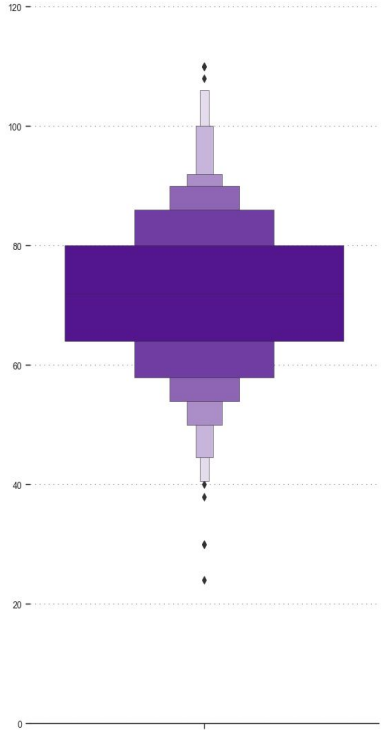
Distribuzione del Glucosio

È evidente come vi sia una correlazione positiva tra il livello di glucosio e la possibilità di essere diagnosticati come diabetici. Allo stesso tempo, è importante notare che esiste un intervallo in cui la presenza del diabete può variare, indicando la necessità di valutare anche altre variabili per una diagnosi accurata.

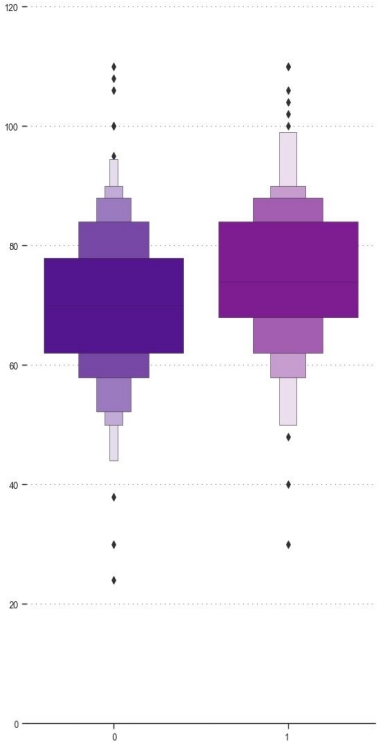


DISTRIBUZIONE DI ALTRE DUE VARIABILI

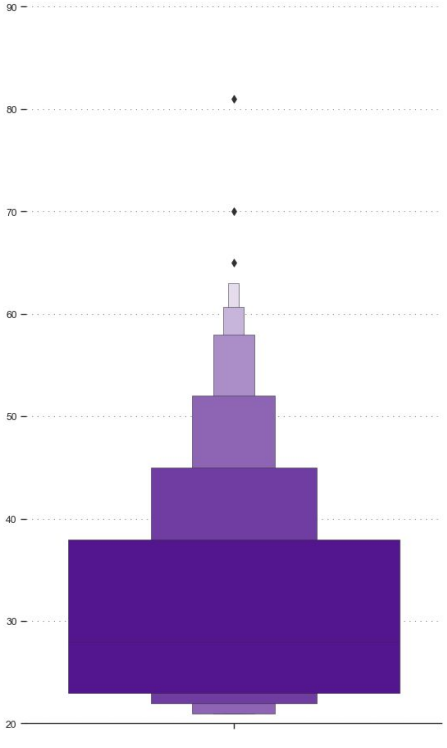
Distribuzione della pressione sanguigna



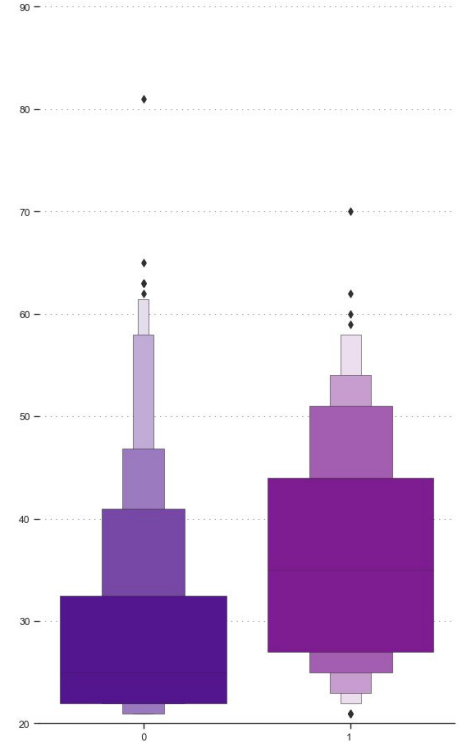
Distribuzione rispetto all'outcome



Distribuzione dell'età



Distribuzione rispetto all'outcome



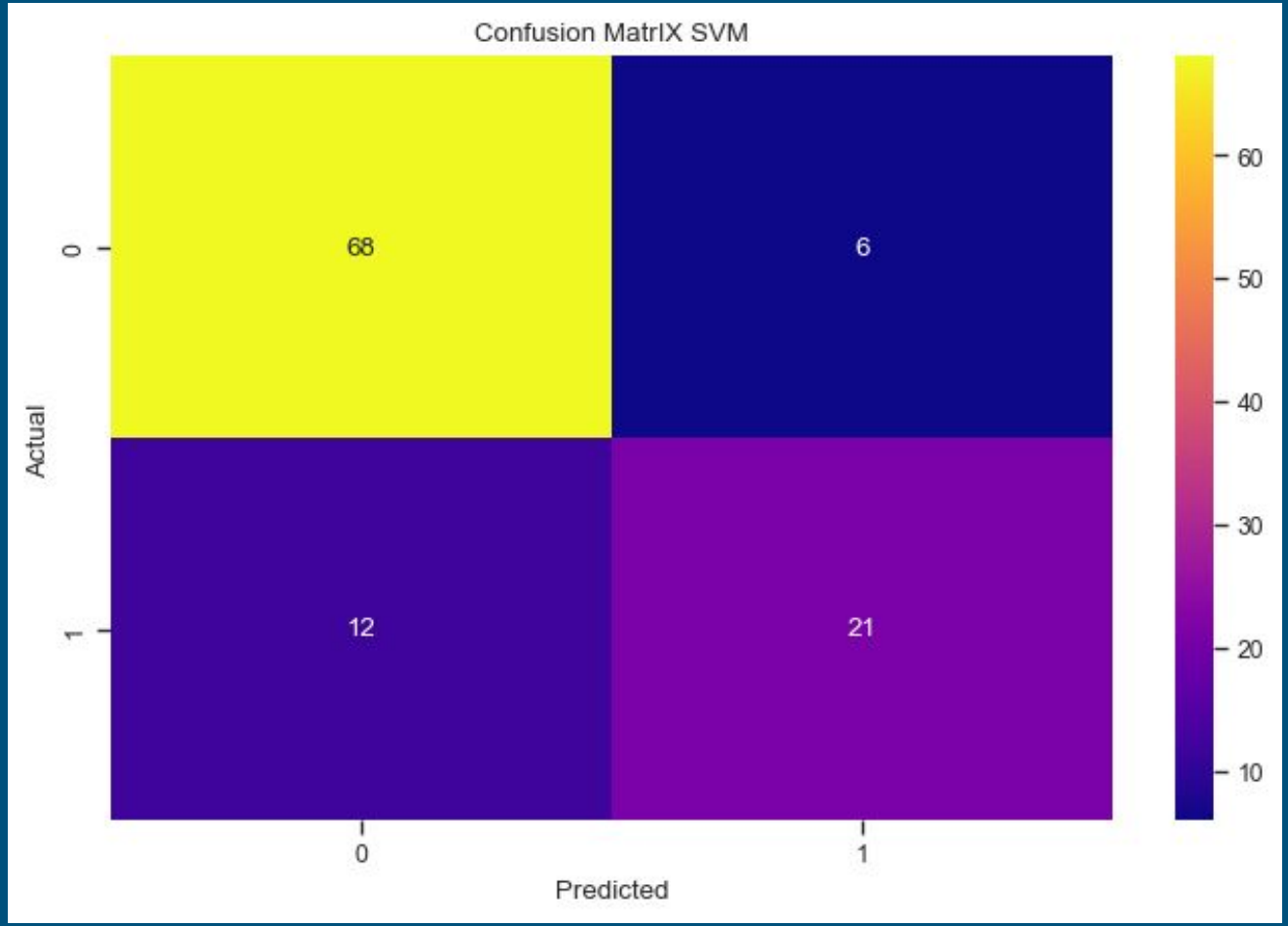
Splitting e addestramento

Il dataset è stato diviso utilizzando lo splitting con una proporzione 80-20, utilizzando l'Outcome come variabile target e tutte le altre colonne come feature. Successivamente, sono stati addestrati due algoritmi di machine learning:
-Support Vector Machine (SVM)

-Regressione Logistica.

SVM

Dall'osservazione della diagonale principale della matrice di confusione, possiamo notare che il modello ha previsto correttamente la maggior parte dei dati. Inoltre, il miscalculation rate è del 16.82%.



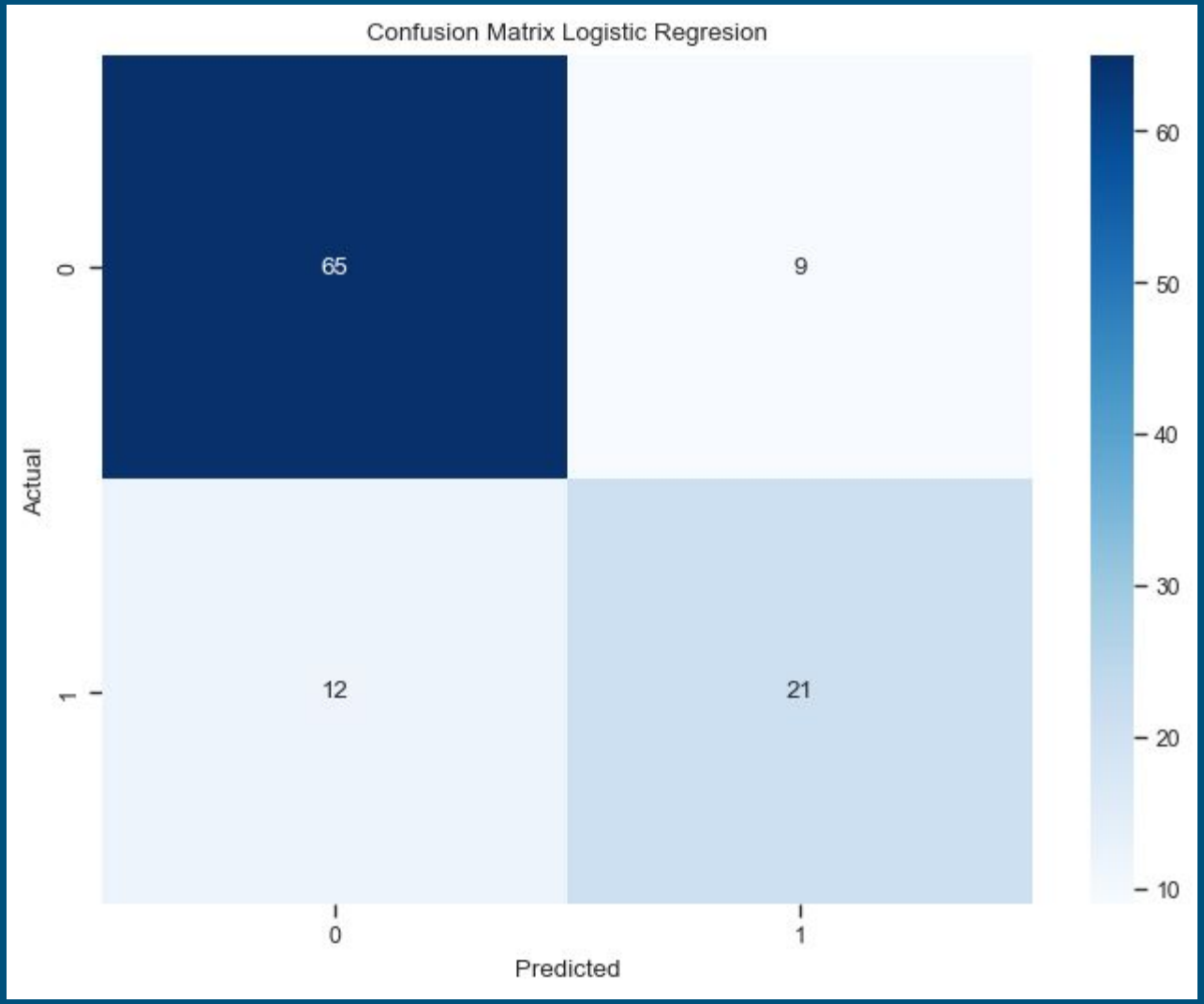
RISULTATI SVM

1)“Precision ” : istanze classificate correttamente come positive rispetto a tutte le istanze classificate come positive (1.00 = tutte le istanze classificate come positive sono effettivamente positive)
2) “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)
3) “F1-score ” rappresenta una media ponderata della “ precisione ” e del “ recall” (1.00 = bilanciamento perfetto tra precisione e recall)
4)“Support” indica il numero di campioni di ogni classe nel set di dati di test

	precision	recall	f1-score	support
0	0.85	0.92	0.88	74
1	0.78	0.64	0.70	33
accuracy			0.83	107
macro avg	0.81	0.78	0.79	107
weighted avg	0.83	0.83	0.83	107

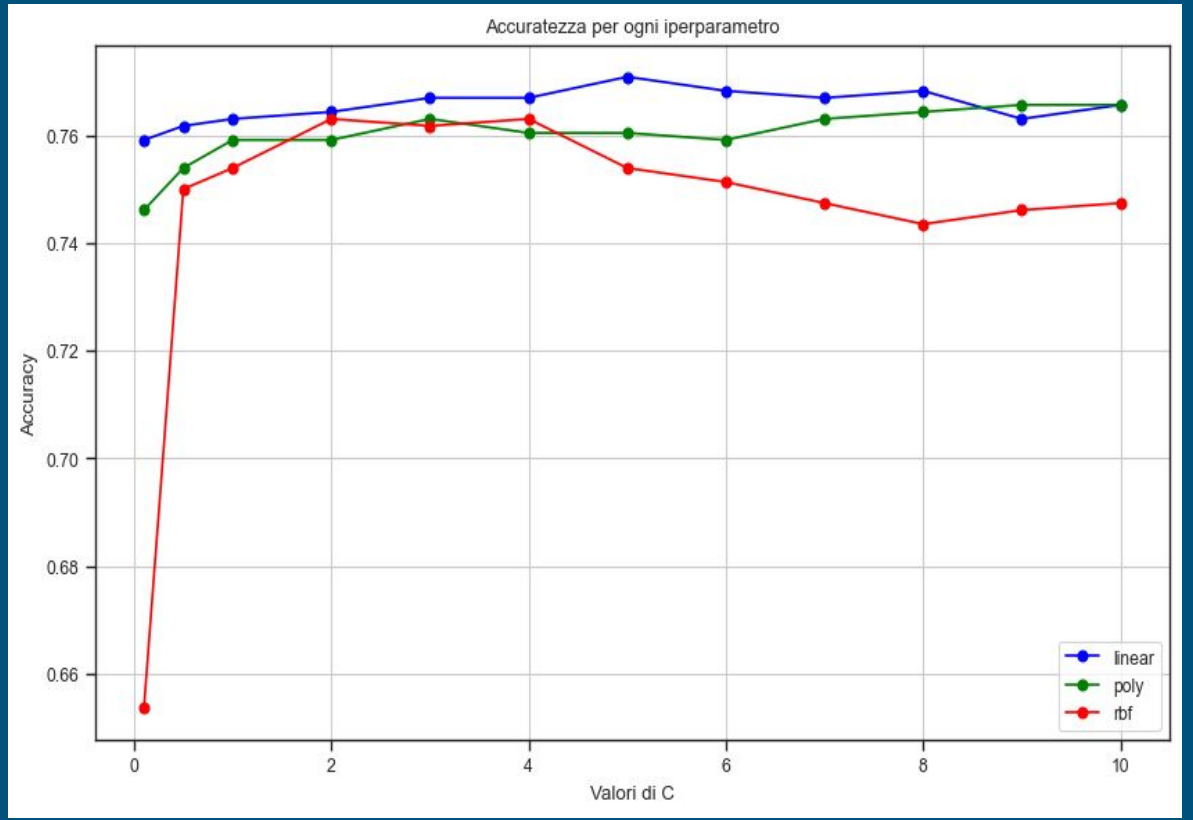
Regressione Logistica

Nella regressione logistica otteniamo un miscalculation rate del 19.63%



Hyperparameter tuning(SVM)

Attraverso il processo di tuning dei parametri, sono stati identificati i migliori iperparametri per il modello. Il set di iperparametri ottimale è risultato essere {'C': 5, 'kernel': 'linear'}.



Boxplot ottimizzati k-13

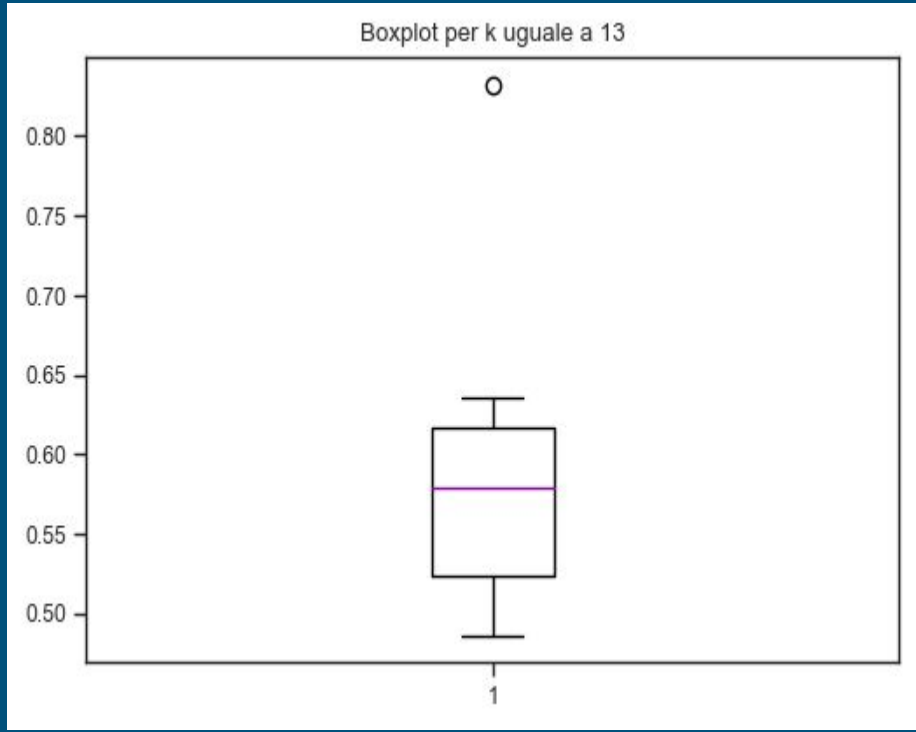
Procediamo con l'addestramento del modello utilizzando gli iperparametri ottimizzati e un valore di k superiore a 10. Per k=13, otteniamo i seguenti risultati:

Media dell'accuratezza:0.5887850467289719

Deviazione standard dell'accuratezza: 0.08415207934722274

Intervallo di confidenza (95%):

(0.5358559902245796, 0.6417141032333642)



k-22

Per k=22:

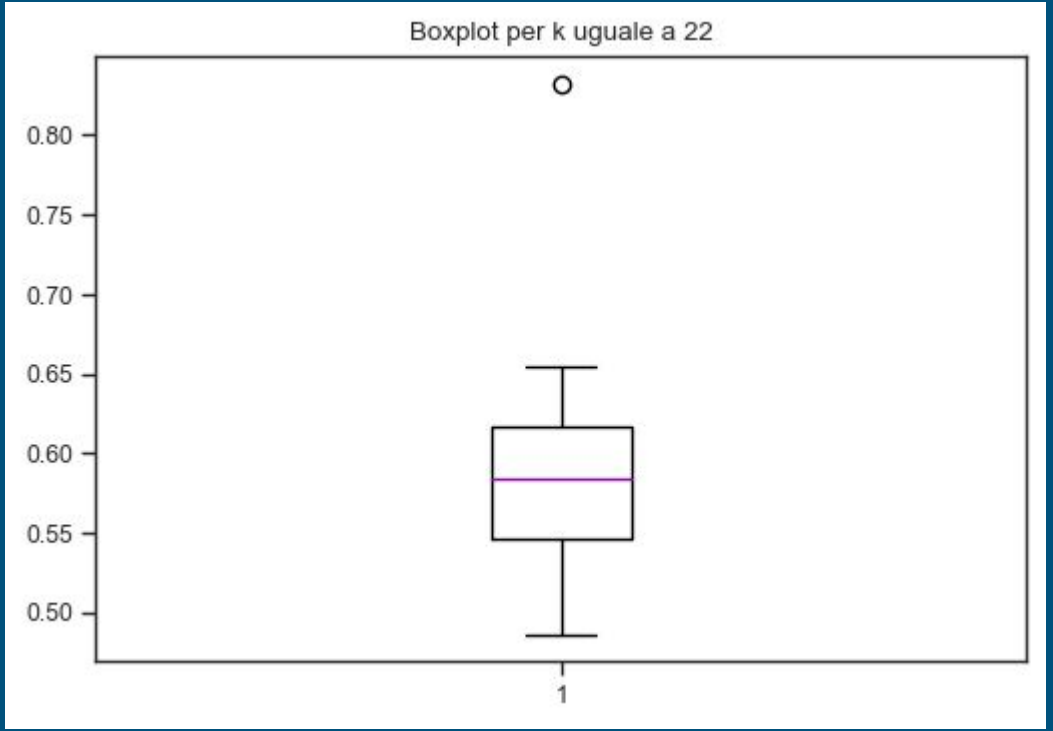
Media dell'accuratezza: 0.5896346644010194

Deviazione standard dell'accuratezza:

0.06947670265304509

Intervallo di confidenza (95%):

(0.5581055152191875, 0.6211638135828513)



k-50

per k=50

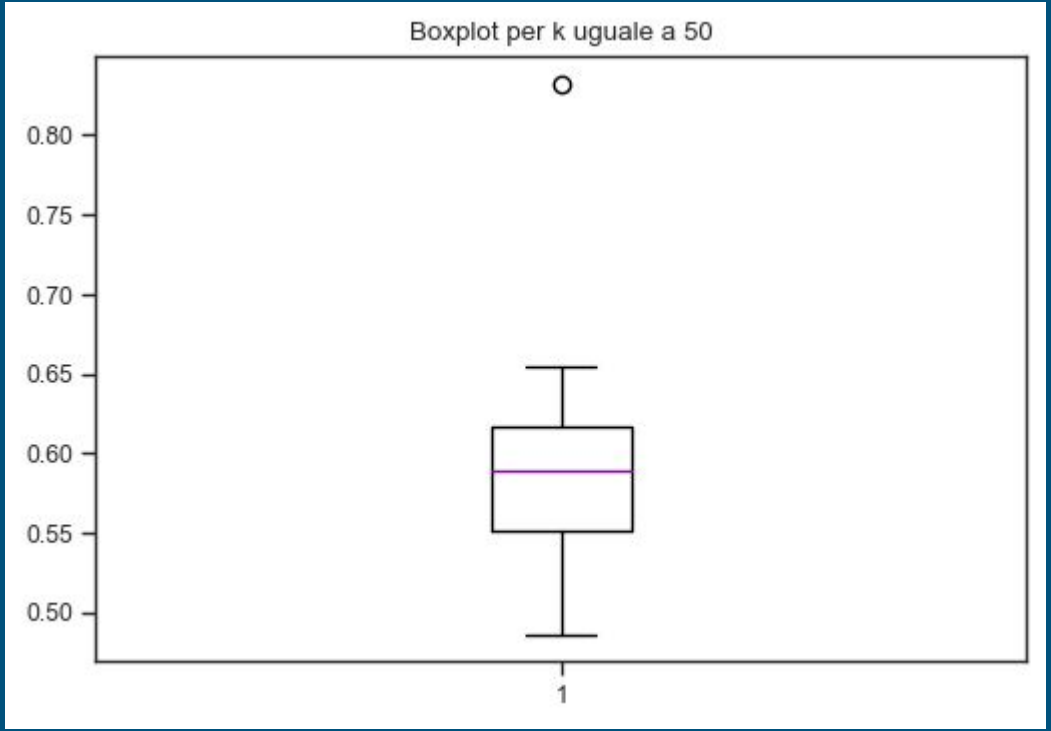
Media dell'accuratezza: 0.5863551401869158

Deviazione standard dell'accuratezza:

0.05332215942574053

Intervallo di confidenza (95%):

(0.5710472986109939, 0.6016629817628377)



FINE