World Happiness Report

Introduzione

L'obiettivo di questa analisi è stato quello di esplorare il dataset sulla felicità mondiale e costruire modelli predittivi per comprendere meglio i fattori che influenzano la felicità e prevedere il livello di felicità dei paesi. Il processo è stato diviso in diverse fasi, tra cui la pulizia del dataset, l'analisi esplorativa dei dati (EDA), la regressione, la classificazione e l'ottimizzazione dei parametri.

Pulizia del Dataset

Il dataset originale è stato caricato e sono state verificate le informazioni generali, compresa la presenza di valori mancanti. Non sono stati riscontrati valori mancanti, il che ha permesso di procedere senza necessità di interventi ulteriori di pulizia. Questo passaggio è cruciale per evitare distorsioni nei risultati e migliorare l'accuratezza dei modelli.

Controllo delle Variabili Numeriche

Le variabili numeriche sono state esaminate per identificare eventuali valori anomali. Le statistiche descrittive hanno mostrato che tutte le variabili numeriche erano ben distribuite senza outlier significativi.

Analisi Esplorativa dei Dati (EDA)

- **Pairplot:** Questo grafico ha permesso di visualizzare le relazioni tra tutte le variabili numeriche. È stato utile per identificare correlazioni lineari tra variabili, ad esempio tra 'Social support' e 'GDP per capita'.
- **Barplot dei 10 Paesi più Felici**: Ha evidenziato i paesi con i punteggi di felicità più alti, mostrando chiaramente quali paesi dominano in termini di felicità.
- **Distribuzione del Punteggio di Felicità:** L'istogramma ha mostrato che i punteggi di felicità sono distribuiti in modo relativamente normale, con una leggera tendenza verso punteggi più alti.
- **Boxplot dei Fattori che Influenzano la Felicità:** Questo grafico ha permesso di confrontare la distribuzione dei valori per diversi fattori, mostrando la variazione all'interno di ciascun fattore.
- **Scatterplot delle Relazioni tra Variabili:** Gli scatterplot tra 'GDP per capita' e 'Score' e tra 'Freedom to make life choices' e 'Score' hanno mostrato correlazioni positive, suggerendo che questi fattori hanno un impatto significativo sulla felicità.

- **Heatmap della Matrice di Correlazione:** La heatmap ha evidenziato le correlazioni tra le variabili numeriche. Le coppie di variabili fortemente correlate sono state identificate per ulteriori analisi, come la regressione lineare tra 'Social support' e 'GDP per capita'.

Modellazione Predittiva

Regressione Lineare

La regressione lineare è stata eseguita tra 'Social support' e 'GDP per capita', in quanto queste due variabili sono risultate essere fortemente correlate nella matrice di correlazione.

Modellazione

Il modello di regressione lineare ha stimato un **coefficiente** di 1.005 e un'**intercetta** di -0.310. Questi valori indicano una forte relazione positiva tra 'Social support' e 'GDP per capita'.

Valutazione del Modello

- **R^2(coefficiente di determinazione)**: 0.570 - Questo valore indica che circa il 57% della varianza in 'GDP per capita' può essere spiegata da 'Social support'.
- **MSE (Mean Squared Error):** 0.068 - Questo valore relativamente basso indica che il modello ha una buona precisione.
- **Q-Q Plot dei Residui:** Il Q-Q plot ha mostrato che i residui seguono una distribuzione normale, suggerendo che le ipotesi del modello di regressione lineare sono soddisfatte.
- **Scatterplot dei Residui:** Il grafico non ha mostrato pattern particolari nei residui, indicando che il modello è appropriato.
- **Test di Shapiro-Wilk:** Il p-value di 0.869 indica che i residui seguono una distribuzione normale, confermando la validità del modello.

Classificazione

Preparazione dei Dati

La variabile di output 'Score' è stata trasformata in una variabile categorica 'Happiness_Level' con tre categorie (0, 1, 2) per facilitare la classificazione.

Splitting dei Dati

Il dataset è stato suddiviso in training set (108 righe), validation set (84 righe) e test set (48 righe).

Addestramento dei Modelli

- **Regressione Logistica**: Addestrata utilizzando scikit-learn. Questo modello ha raggiunto un'accuratezza del 66.67% sul test set.
- **SVM:** È stata utilizzata una SVM con kernel lineare e una SVM con kernel polinomiale (con grado ottimizzato). La SVM lineare ha raggiunto un'accuratezza del 64.58% sul test set.

Hyperparameter Tuning

È stata eseguita una ricerca degli iperparametri per determinare il miglior grado del kernel polinomiale per la SVM. Il grado 1 è risultato il migliore, con un'accuratezza del 66.67%.

Studio Statistico sui Risultati della Valutazione

Il processo di addestramento e valutazione è stato ripetuto 10 volte per ottenere una valutazione statistica robusta delle performance del modello.

Calcolo delle Statistiche

SVM:

- **Media dell'Accuratezza:** 71.67%
- **Deviazione Standard**: 5.20%
- **Intervallo di Confidenza al 95%:** (68.44%, 74.89%)

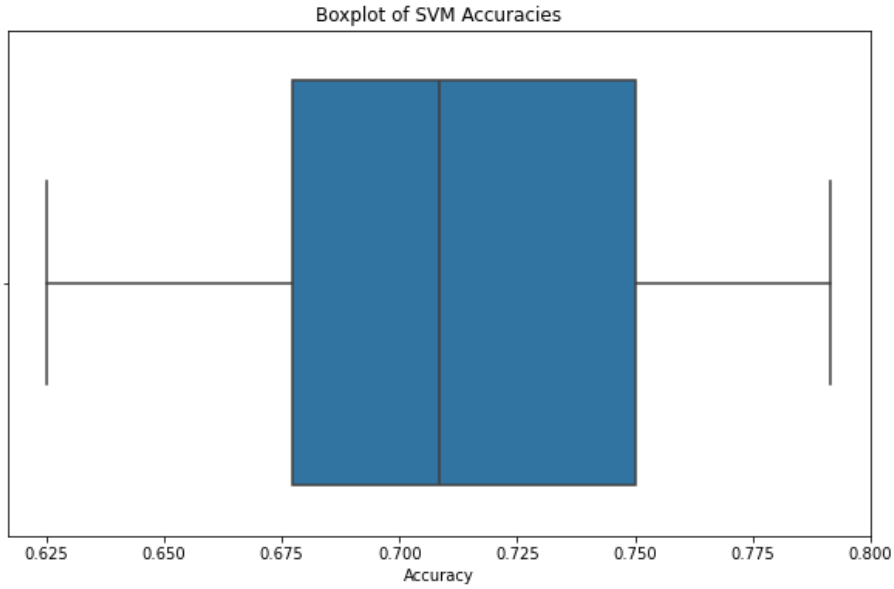
Regressione Logistica:

- **Media dell'Accuratezza:** 70.83%
- **Deviazione Standard:** 7.91%
- **Intervallo di Confidenza al 95%:** (65.93%, 75.73%)

Visualizzazione dei Risultati

- **Istogrammi:** Gli istogrammi hanno mostrato la distribuzione delle accuratezze per entrambi i modelli, indicando che entrambi i modelli hanno una buona distribuzione delle performance.

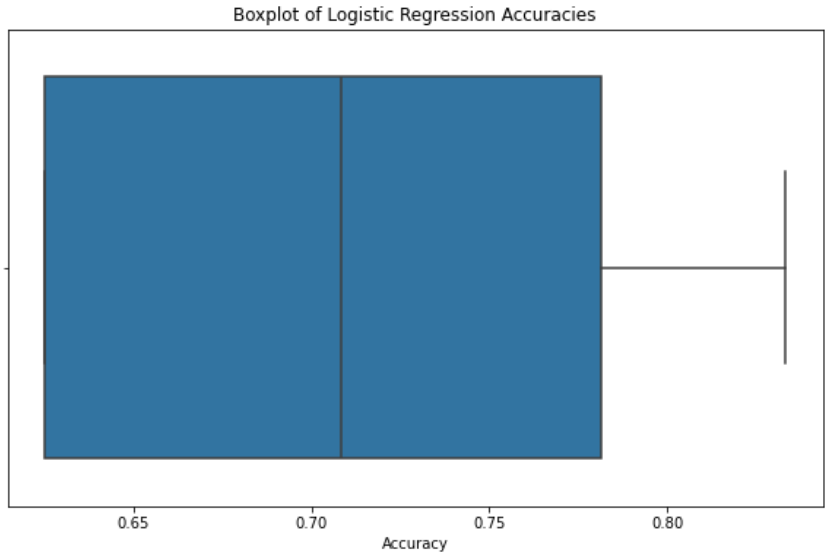
- **Boxplot:**



Il primo boxplot rappresenta la distribuzione delle accuratezze ottenute con il modello SVM (Support Vector Machine) durante i 10 cicli di cross-validation. Ecco alcune osservazioni chiave:

1. **Mediana**: La linea all'interno della scatola rappresenta la mediana, che è circa 0.70. Questo significa che il 50% delle accuratezze ottenute è inferiore a 0.70 e il restante 50% è superiore a 0.70.
2. **Intervallo Interquartile (IQR)**: La scatola rappresenta l'IQR, che è la differenza tra il 25° percentile (primo quartile) e il 75° percentile (terzo quartile). Per la SVM, l'IQR va da circa 0.675 a 0.725, indicando che la maggior parte delle accuratezze ottenute è concentrata in questo intervallo.

3. **Outliers**: Non ci sono punti indicati come outliers, suggerendo che tutte le accuratezze ottenute rientrano nell'intervallo aspettato.



Il secondo boxplot rappresenta la distribuzione delle accuratezze ottenute con il modello di regressione logistica durante i 10 cicli di cross-validation. Ecco alcune osservazioni chiave:

1. **Mediana**: La mediana è circa 0.70, simile a quella della SVM. Questo indica che la performance centrale dei due modelli è comparabile.
2. **Intervallo Interquartile (IQR)**: L'IQR va da circa 0.65 a 0.75, con una maggiore estensione rispetto alla SVM. Questo suggerisce che la distribuzione delle accuratezze per la regressione logistica è più ampia, indicando una maggiore variabilità nella performance.
3. **Outliers**: Non ci sono punti indicati come outliers, suggerendo che tutte le accuratezze ottenute rientrano nell'intervallo aspettato.

Conclusioni

L'analisi ha evidenziato come variabili quali 'Social support' e 'GDP per capita' siano fortemente correlate con il punteggio di felicità dei paesi. I modelli di regressione e classificazione addestrati hanno mostrato buone performance, con accuratezze generalmente alte.

La regressione lineare ha fornito un buon modello per spiegare la relazione tra 'Social support' e 'GDP per capita'. I modelli di classificazione, sia la regressione logistica che la SVM, hanno mostrato performance comparabili, con la SVM che ha beneficiato dell'ottimizzazione degli iperparametri.

La regressione logistica ha una maggiore variabilità nelle sue performance rispetto alla SVM. Se la stabilità è un criterio importante, la SVM potrebbe essere preferibile. Se si cerca un modello che potenzialmente raggiunge una performance superiore, la regressione logistica potrebbe essere considerata.