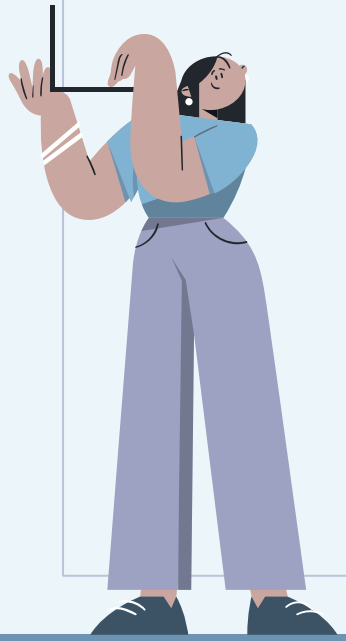


Analisi Predittiva del Successo Accademico: Un Approccio di Machine Learning

Tesi a cura di Simone Magli

Passaggi chiave del progetto

01 Pre-Processing

02 Exploratory Data Analysis

03 Splitting

04 Predizione di una variabile target

05 Hyperparameter Tuning

06 Conclusioni

01 - Pre-processing

La fase di **pre-processing** è il primo passaggio che viene effettuato all'interno del progetto e consiste nell'ispezione e preparazione del dataset per i futuri passaggi:

- Gestione dei valori mancanti
- Rimozione delle variabili non utili
- Codifica delle variabili categoriche

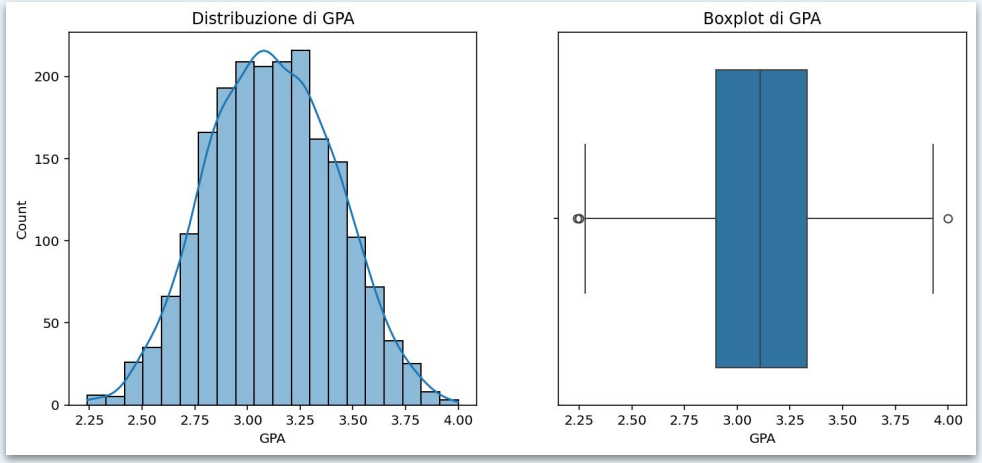
Student_ID	Study_Hou...	Extracurri...	Sleep_Hou...	Social_Ho...	Physical_A...	GPA	Stress_Level
1	6.9	3.8	8.7	2.8	1.8	2.99	Moderate
2	5.3	3.5	8	4.2	3	2.75	Low
3	5.1	3.9	9.2	1.2	4.6	2.67	Low
4	6.5	2.1	7.2	1.7	6.5	2.88	Moderate
5	8.1	0.6	6.5	2.2	6.6	3.51	High
6	6	2.1	8	0.3	7.6	2.85	Moderate
7	8	0.7	5.3	5.7	4.3	3.08	High
8	8.4	1.8	5.6	3	5.2	3.2	High
9	5.2	3.6	6.3	4	4.9	2.82	Low
10	7.7	0.7	9.8	4.5	1.3	2.76	Moderate
11	9.7	3.6	8	2.5	0.2	3.43	High
12	6.9	1.1	9.1	2.1	4.8	2.97	Moderate
13	6.4	2.2	5.7	4.8	4.9	2.82	High
14	5	3.3	8.5	4.4	2.8	2.87	Low
15	8.9	0.3	6.8	0.7	7.3	3.4	High
16	6.7	0.3	6.6	2	8.4	3.2	Moderate

02 - Exploratory Data Analysis

Che cos'è l'EDA?

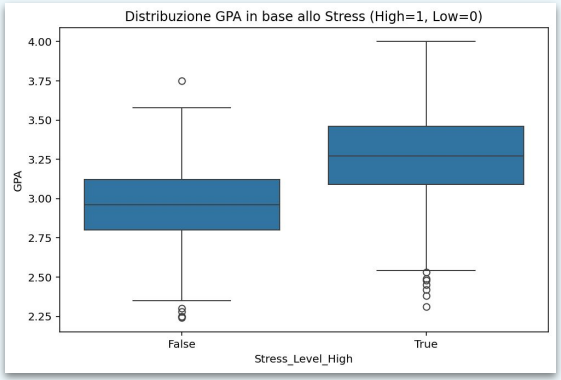
E' il processo che prevede l'utilizzo di **riepiloghi numerici** e **visualizzazioni** per studiare i dati e identificare potenziali **relazioni** tra le variabili.
Sono stati usati principalmente quattro tipi di visualizzazione dei dati diversi.

1) Distribuzioni Univariate

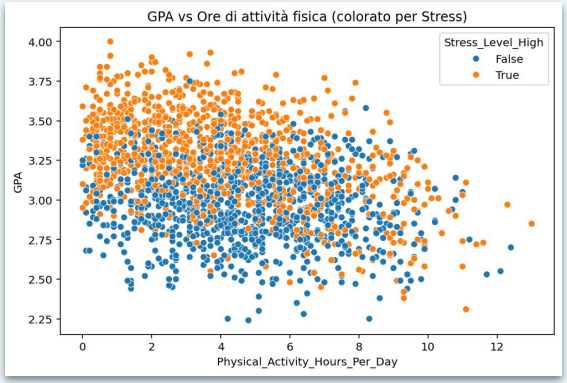


02 - Exploratory Data Analysis

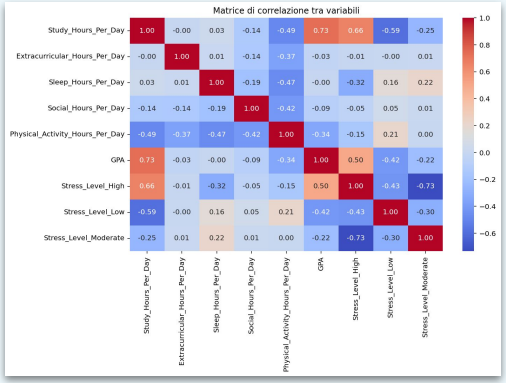
2) Distribuzioni Bivariate



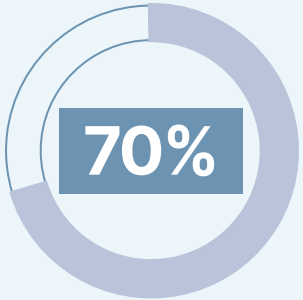
3) Distribuzioni Multivariate



4) Matrici di Correlazione

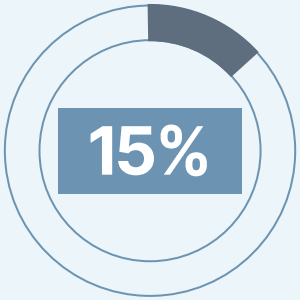


03 - Splitting



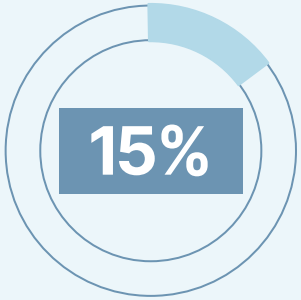
Training set

Utilizzato per **addestrare** i modelli e per **l'ottimizzazione** degli iperparametri (tuning).



Validation set

Impiegato per il **confronto** delle prestazioni e la **selezione** preliminare tra i modelli.



Test set

Riservato per ottenere una **stima** conclusiva, non contaminata, delle predizioni su dati mai visti.

04 - Predizione di una variabile target

Metriche scalari di valutazione

R^2

Valore che indica quanto il modello di predizione utilizzato è in grado di **spiegare la varianza** dei dati.
Idealmente ha una valutazione che è compresa tra 0 e 1.
Quando il valore è **0** significa che il modello è **inutile**, di conseguenza quando è vicino ad **1** significa che il modello spiega molto **bene** la varianza.

MSE

Indicatore che va a quantificare gli **errori** concreti del modello utilizzato. Calcola la differenza tra il valore predetto dal modello e il valore reale (che poi viene elevato al quadrato per eliminare i segni negativi). Al contrario della metrica precedente più i valori sono **bassi** più significa che il modello **non commette errori.**

04 - Regressione Lineare Multivariata

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon$$

Migliore Adattamento Lineare

Cerca la **"linea retta" ottimale** che minimizza l'errore per tutti i punti dati (Minimi Quadrati).

Alta Interpretabilità

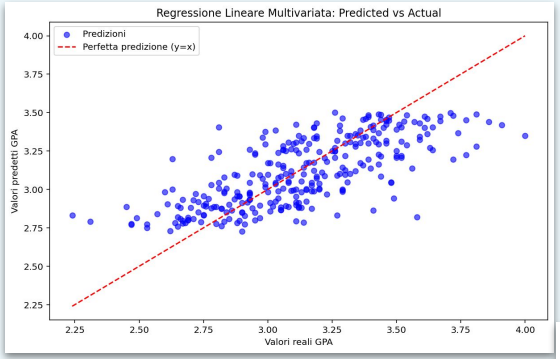
I coefficienti **β** forniscono la **misura diretta** dell'impatto di ogni variabile.

Sensibile a Variazioni

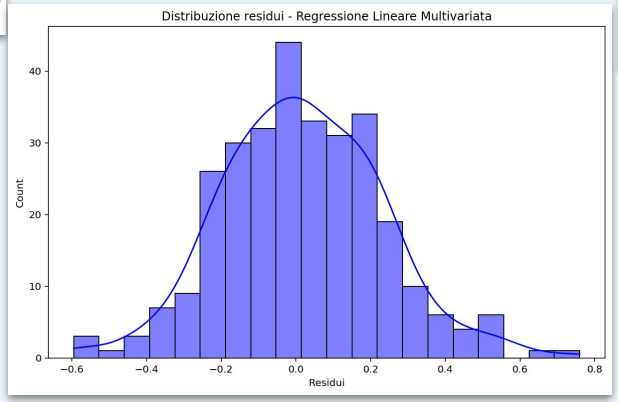
Molto suscettibile agli **outlier** e alla **multicollinearità** dei dati.

04 - Regressione Lineare Multivariata

Predicted vs Actual



Distribuzione Residui



Metriche di Validazione

```
=== Regressione Lineare Multivariata ===
Coefficiente/i:
  Study_Hours_Per_Day: 0.1248
  Extracurricular_Hours_Per_Day: -0.0412
  Sleep_Hours_Per_Day: -0.0287
  Social_Hours_Per_Day: -0.0269
  Physical_Activity_Hours_Per_Day: -0.0280
  Stress_Level_High: 0.0080
  Stress_Level_Low: 0.0117
  Stress_Level_Moderate: -0.0197
Intercetta: 2.6750
R^2 (Validation): 0.5251
MSE (Validation): 0.0462
```

04 - Random Forest Regressor

$$Y(x) = 1/T * \sum_{t=1}^{T} ht(x)$$

Modello Ensemble

La previsione è la **media** delle previsioni di T alberi decisionali indipendenti.

Robustezza e Non-Linearità

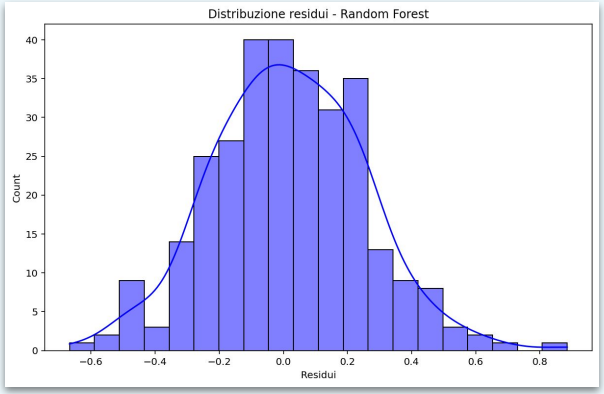
Cattura **relazioni complesse** ed è molto robusto agli outlier.

Bassa Trasparenza

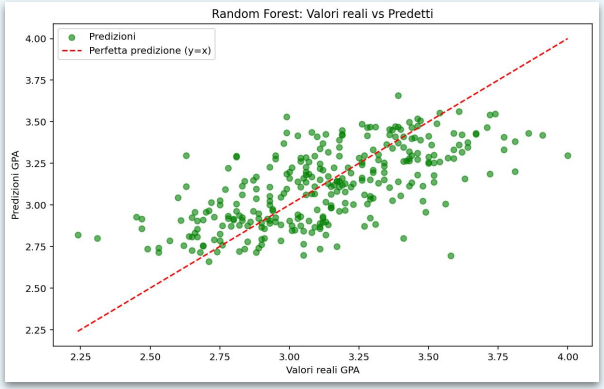
Il meccanismo interno di decisione è **complesso e poco trasparente**.

04 - Random Forest Regressor

Distribuzione Residui



Predicted vs Actual

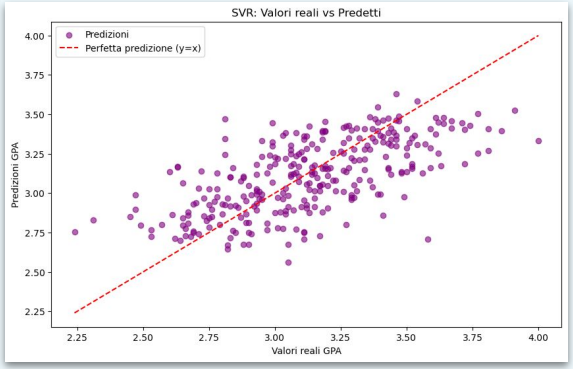


Metriche di Validazione

```
=== Random Forest Regressor ===
R^2: 0.4210
MSE: 0.0564
```

04 - Support Vector Regression (SVR)

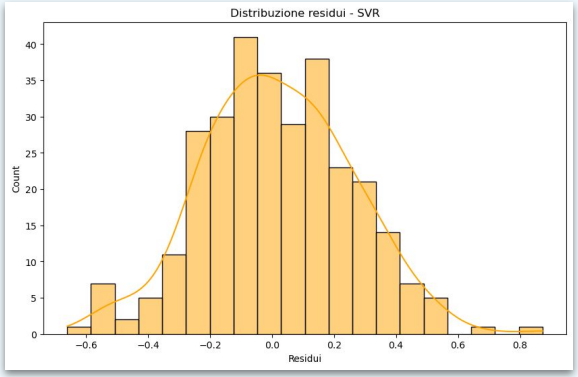
Predicted vs Actual



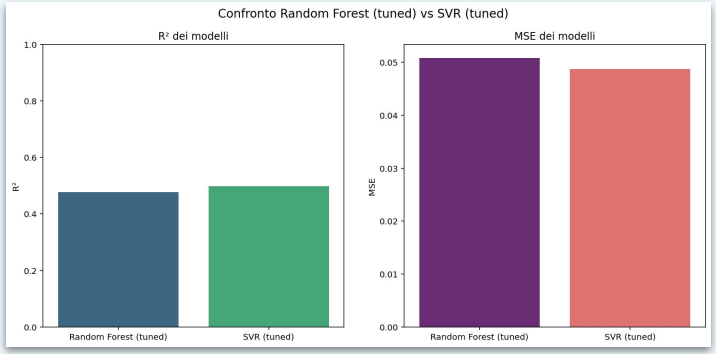
Metriche di Validazione

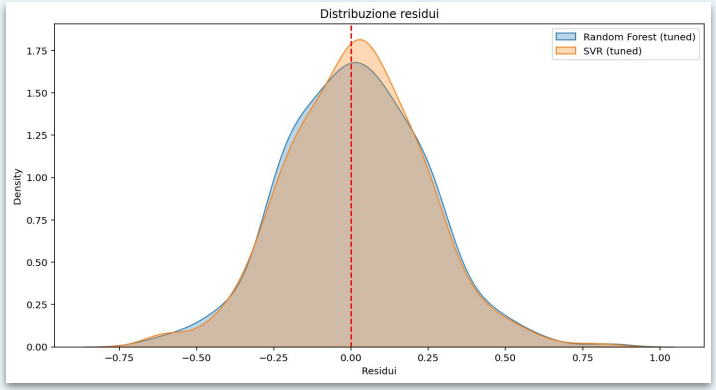
```
=== Support Vector Regression (RBF Kernel) ===
R^2: 0.4034
MSE: 0.0581
```

Distribuzione Residui



05 - Hyperparameter Tuning





Processo che mira a trovare la **miglior combinazione** di iperparametri per poter sfruttare i modelli di predizione nella loro miglior forma possibile, attraverso il metodo della **Grid Search.**

Risultati delle metriche:

- Random Forest Regressor:
  - R^2: 0.4778
  - MSE: 0.0508
- SVR:
  - R^2: 0.4988
  - MSE: 0.0488

06 - Conclusioni

Modello	R^2	MSE
Regressione Lineare Multipla	0.5251	0.0462
Random Forest Regressor	0.4210	0.0564
Support Vector Regression	0.4034	0.0581
Random Forest - tuned	0.4778	0.0508
SVR - tuned	0.4988	0.0488

06 - Conclusioni

In conclusione sono stati effettuati due passaggi per **decretare** il modello migliore:

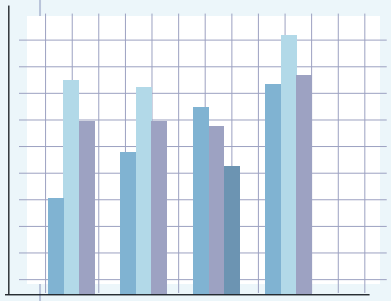
1. I tre migliori modelli della tabella sono stati eseguiti sul **test set**.
2. I due migliori modelli trovati dall'analisi dei risultati sul test set sono stati confrontati attraverso un'**analisi statistica** (k ripetizioni, calcolo delle metriche di valutazione e intervalli di confidenza al 95%)

```
=== VALUTAZIONE FINALE SU TEST SET ===

Linear Regression (Multivariata)
  R²  = 0.5317
  MSE = 0.0391

Random Forest (tuned)
  R²  = 0.5073
  MSE = 0.0412

SVR (tuned)
  R²  = 0.5302
  MSE = 0.0393
```

```
=== Linear Regression ===
R² medio: 0.5288 ± 0.0186
IC 95% R²: (np.float64(0.5217510833775572), np.float64(0.5359068427492231))
MSE medio: 0.0415 ± 0.0019
IC 95% MSE: (np.float64(0.04077330376950911), np.float64(0.042182891313184874))

=== SVR (tuned) ===
R² medio: 0.5146 ± 0.0192
IC 95% R²: (np.float64(0.5072626093359199), np.float64(0.5218711645460964))
MSE medio: 0.0427 ± 0.0021
IC 95% MSE: (np.float64(0.041944374671972935), np.float64(0.043541048059278366))
```

Fine

Grazie per l’attenzione!



Simone Magli - Informatica per il management
0001069295