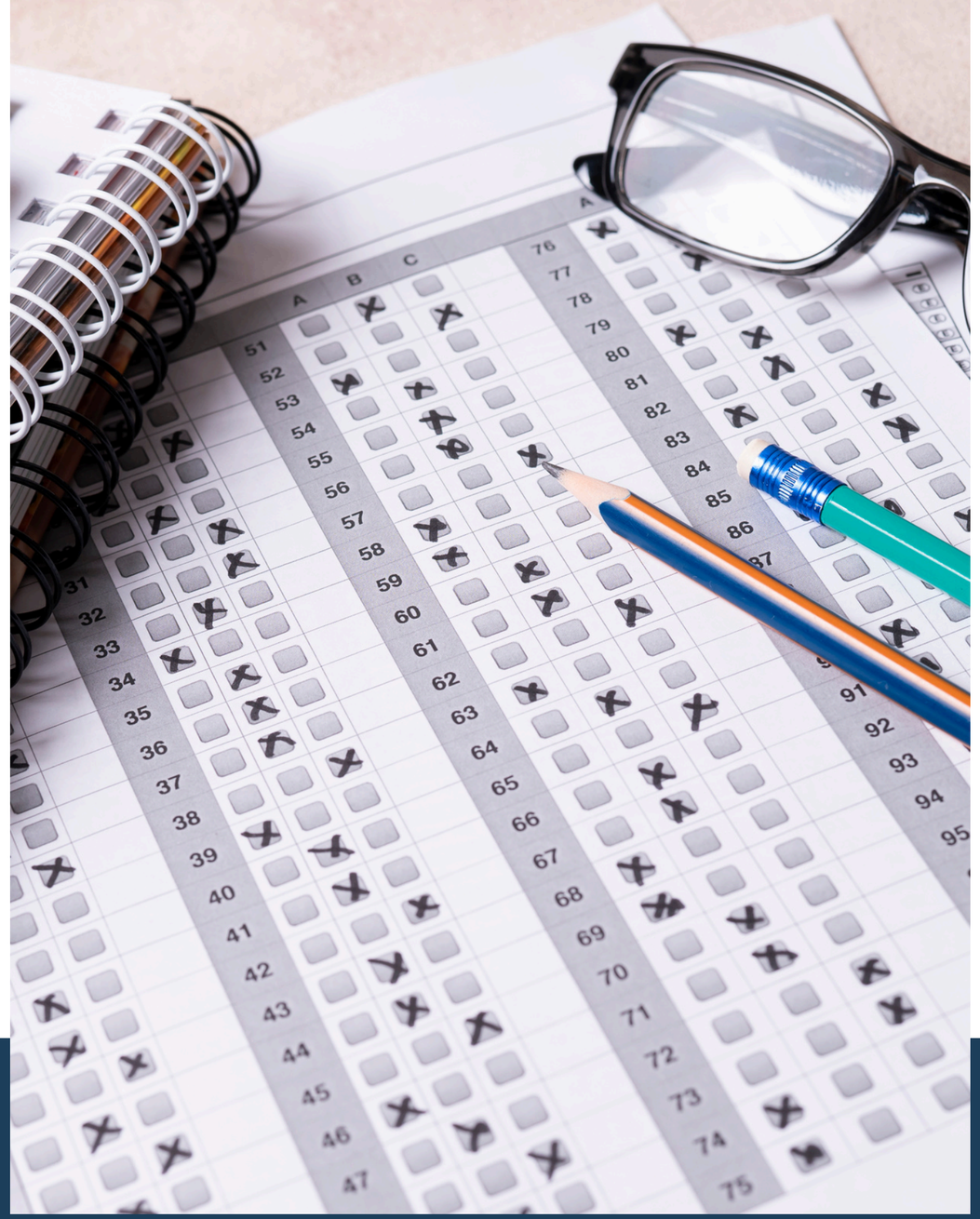
Dai Voti ai Dati

Analisi e Predizione dei Risultati Scolastici per Identificare i Fattori Chiave e Predirne i Risultati

Alessia Profeta

alessia.profeta@gmail.com

Indice



Overview

Il presente progetto si propone di analizzare le performance scolastiche di un campione di studenti delle scuole superiori, con l'obiettivo di identificare i fattori che influenzano maggiormente i risultati finali e di effettuare valutazioni predittive.

Risorse esterne

Datasource

Questo lavoro si basa sul dataset por, che raccoglie dati sulle performance scolastiche, nonché variabili personali e socio-demografiche degli studenti delle scuole superiori, consentendo un'analisi approfondita dei diversi aspetti che influenzano i risultati accademici degli studenti.

GitHub

Il repository contiene tutto il codice sorgente (Alchol_Youth_EDA_PDA.ipynb) utilizzato per l'analisi dei dati e la costruzione dei modelli predittivi e contiene tutte le spiegazioni aggiuntive e i grafici non presenti in queste slide.

Fasi di sviluppo

01

Analisi Esplorativa

Sarà eseguito un **EDA** al fine di individuare i fattori che influenzano maggiormente i risultati scolastici.

02

Analisi Predittiva

Mediante tecniche di **ML**, si costruiranno modelli predittivi al fine di stimare i voti finali degli studenti

Analisi Esplorativa

01

Analisi del voto

Esamina la distribuzione e le caratteristiche dei voti.

02

Correlazione col voto

Come il voto finale sia correlato con altre variabili

03

Analisi delle feature

L'analisi della relazione del voto con le altre variabili del dataset

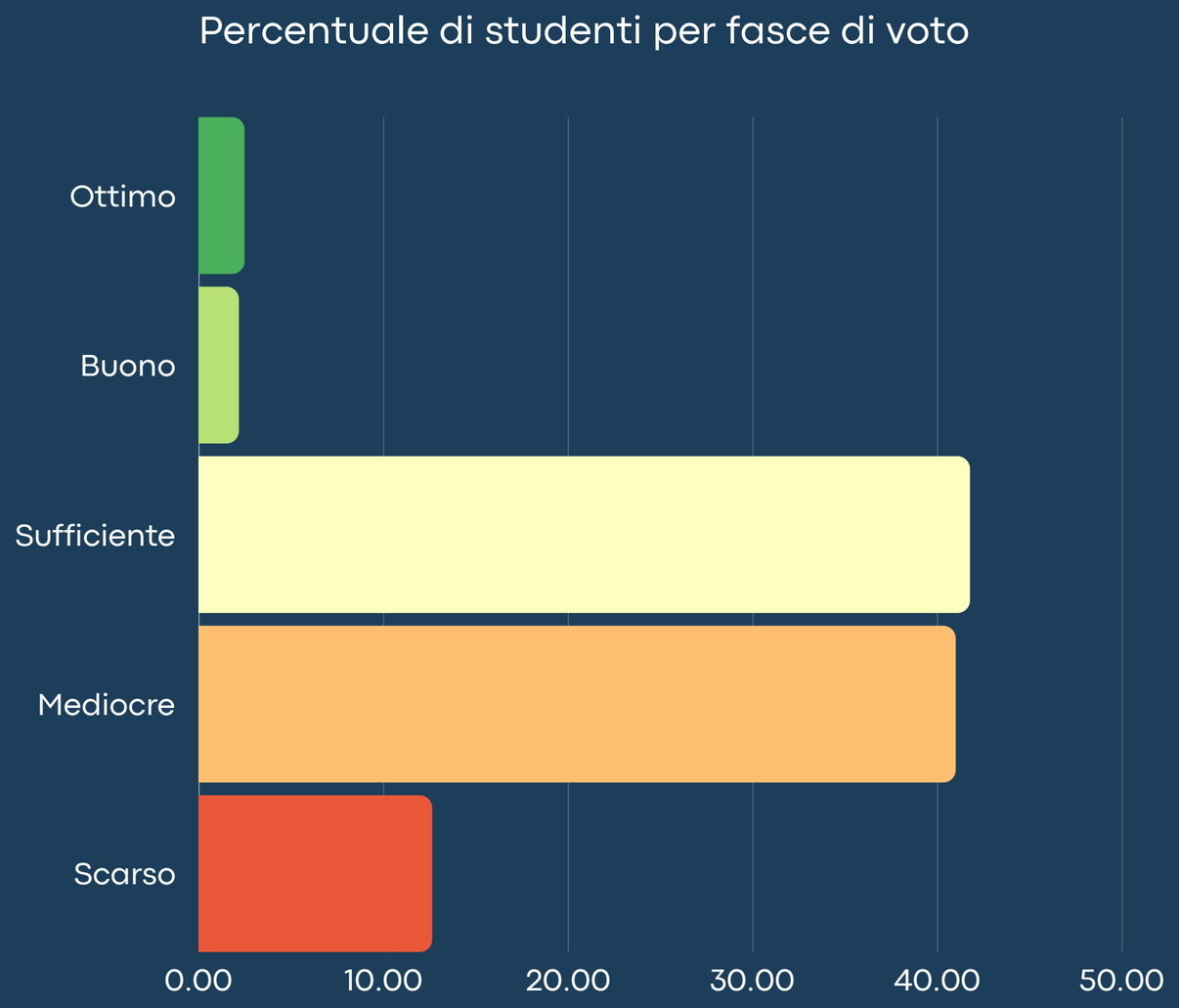
Analisi del voto

Il Voto finale è espresso in una scala numerica [0, 19] e la **media totale** è:

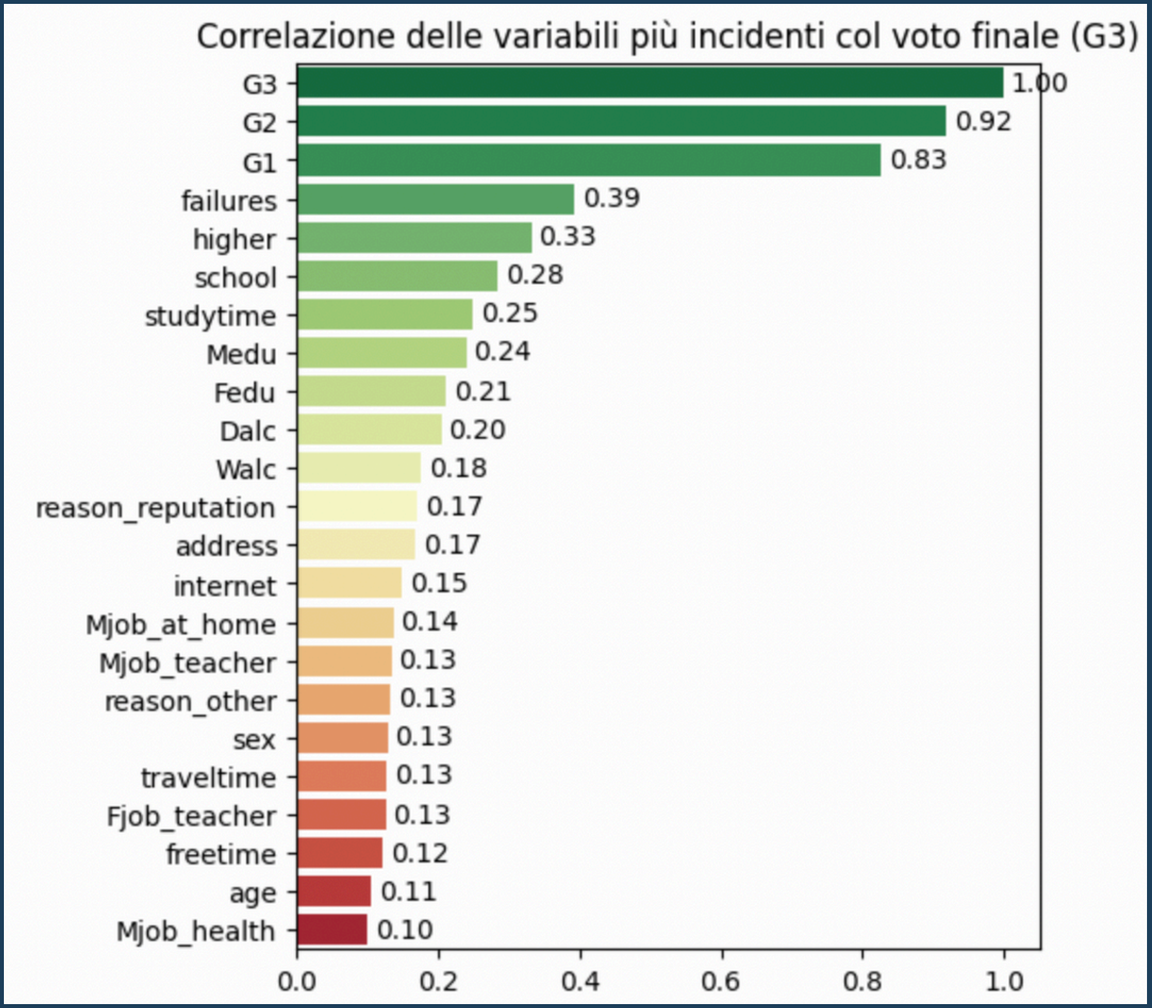
11.91

Tramite un processo di Data Binning, il voto finale è stato trasformato in variabile categorica-ordinale per migliorarne la leggibilità tramite la seguente scala:

scarso: [0, 3]
mediocre: [4, 7]
sufficiente: [8, 11]
buono: [12, 15]
ottimo: [16, 19]



Correlazione col voto finale



Per comprendere meglio le relazioni tra il voto finale e le altre variabili, è stata costruita una matrice di correlazione.

Il grafico associato, riportato a destra, illustra visivamente la correlazione tra il voto finale (G3) e le altre variabili.

Analisi delle principali influenze

Aree di Studio Analizzate per Identificare le Principali Influenze sul Voto Finale

Dati Anagrafici

Istituti Scolastici

Ambiente Familiare

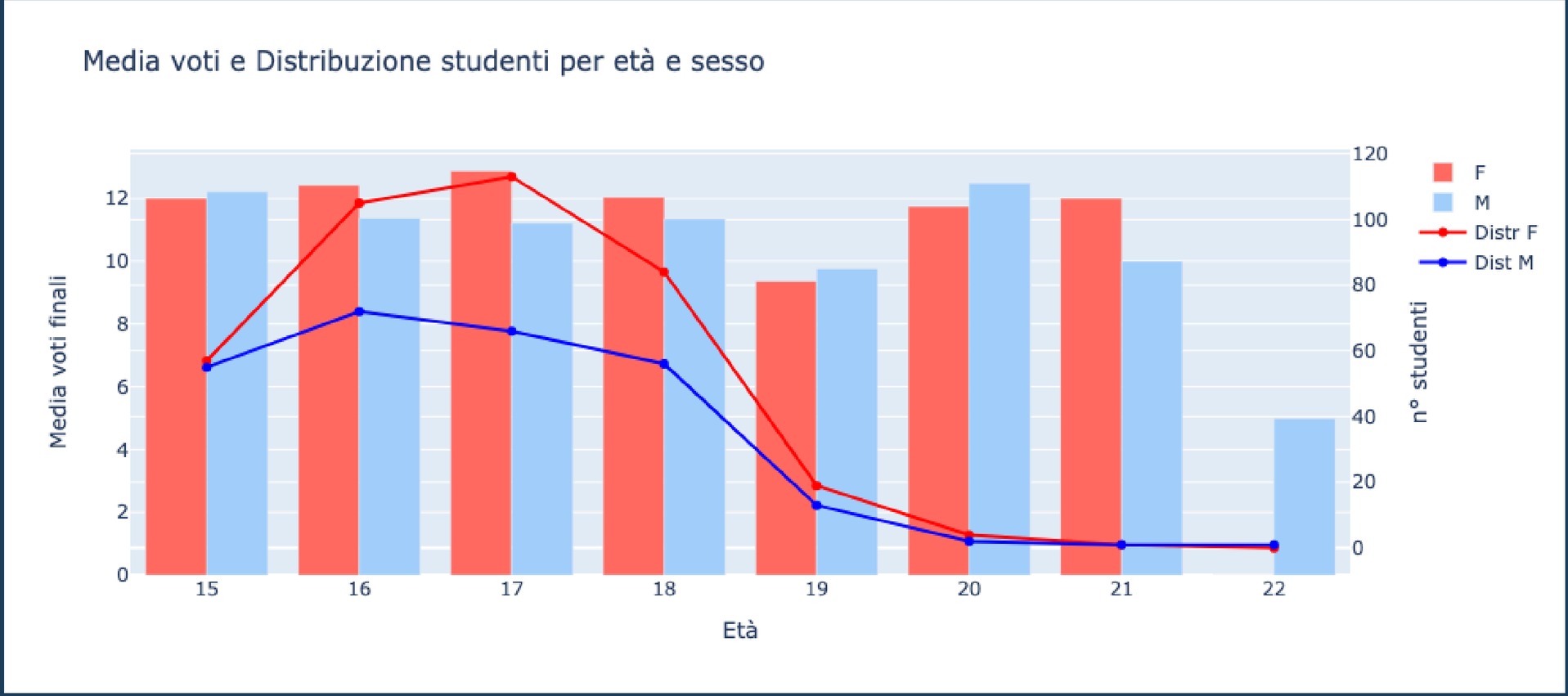
Gestione del Tempo

Consumo Alcolici

Ulteriori Analisi

Dati Anagrafici

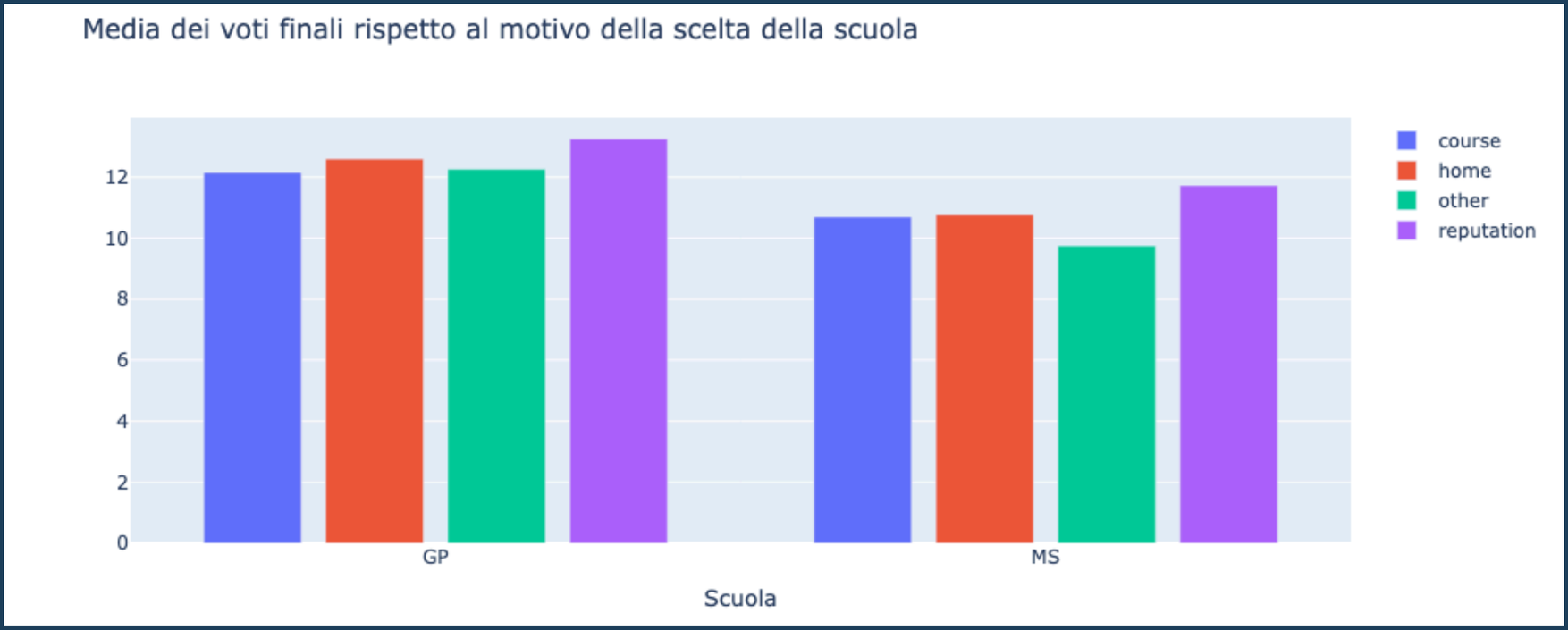
Il grafico mostra la media dei voti finali in relazione all'età e al sesso. Le linee indicano come i voti medi variano per ciascun gruppo di età, suddivisi tra maschi e femmine. Il numero di studenti per età e sesso (grafico a linee) è incluso per evidenziare che le medie non sono assolute ma condizionate dalla distribuzione degli studenti.



Istituti Scolastici

In questa sezione, presentiamo un confronto tra istituti ('GP' - Gabriel Pereira o'MS' - Mousinho da Silveira) e rendimento scolastico.
Nella tabella sono riassunte, per ogni istituto la media dei voti: degli studenti, degli studenti bocciati e di quelli che intendono proseguire gli studi.
Questo ci fornisce una panoramica iniziale delle differenze tra le scuole in termini di successo accademico e aspirazioni future degli studenti.

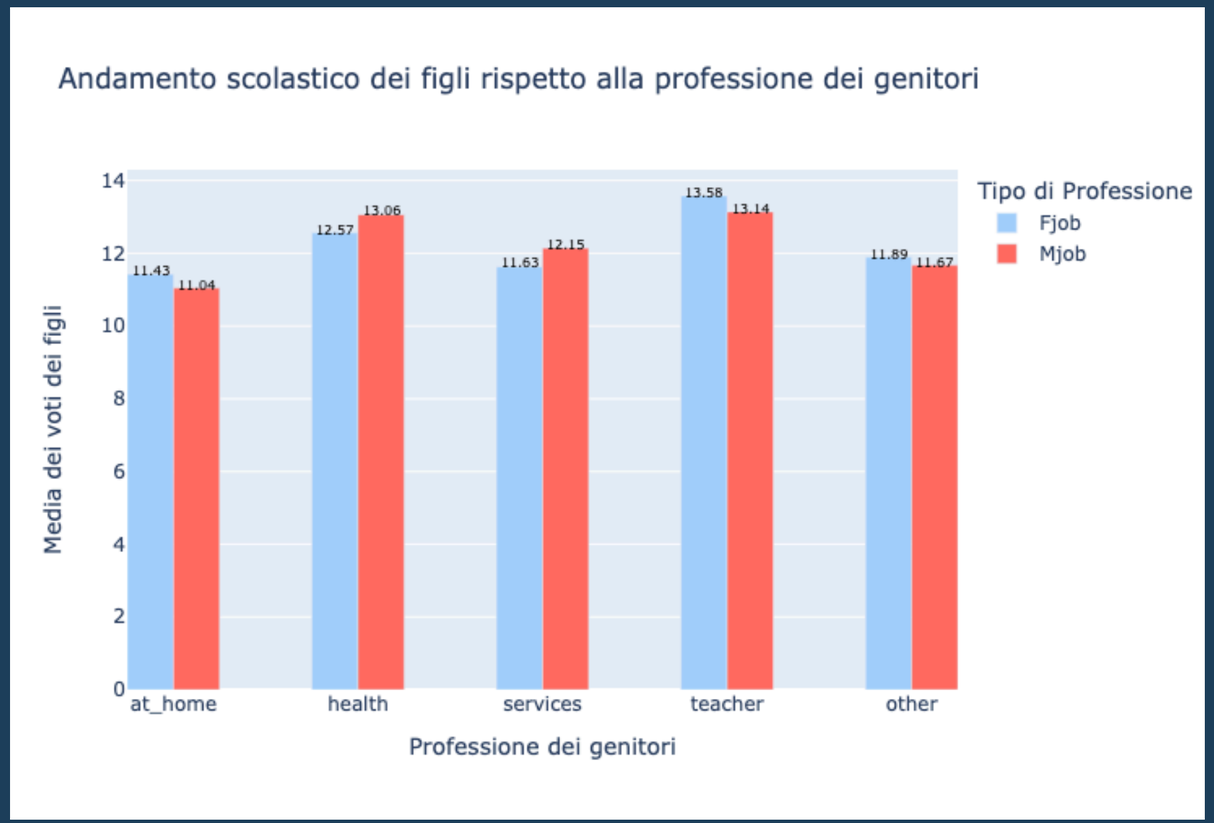
sch	avg_sch	fail_per	G3_higher
GP	12.58	4.26	12.86
MS	10.65	5.31	11.06

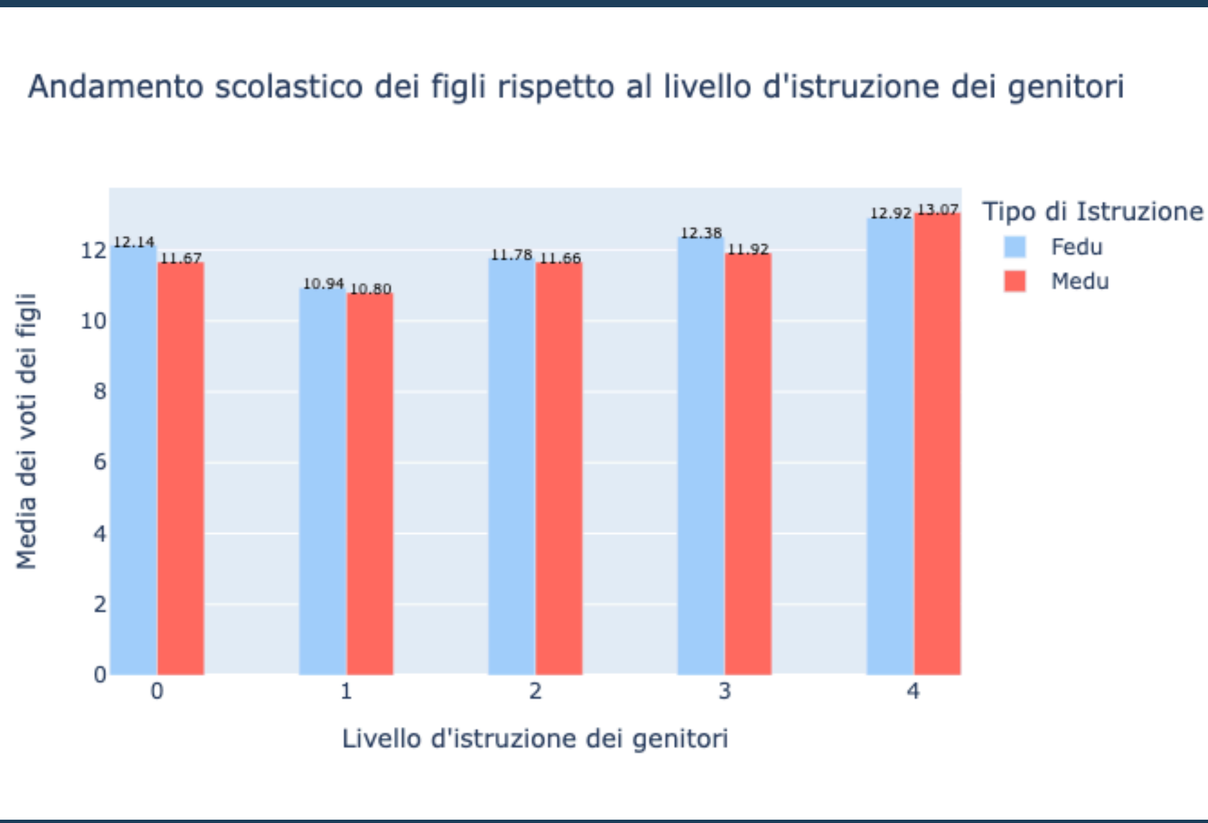


GP: Gli studenti che frequentano GP tendono a ottenere voti finali più alti rispetto a quelli di MS. Inoltre, la percentuale di fallimenti è inferiore, suggerendo un tasso di successo più elevato.
MS: Gli studenti di MS mostrano una media di voti inferiore e una percentuale di fallimenti più alta. Anche il voto finale per coloro che proseguono gli studi è relativamente più basso.

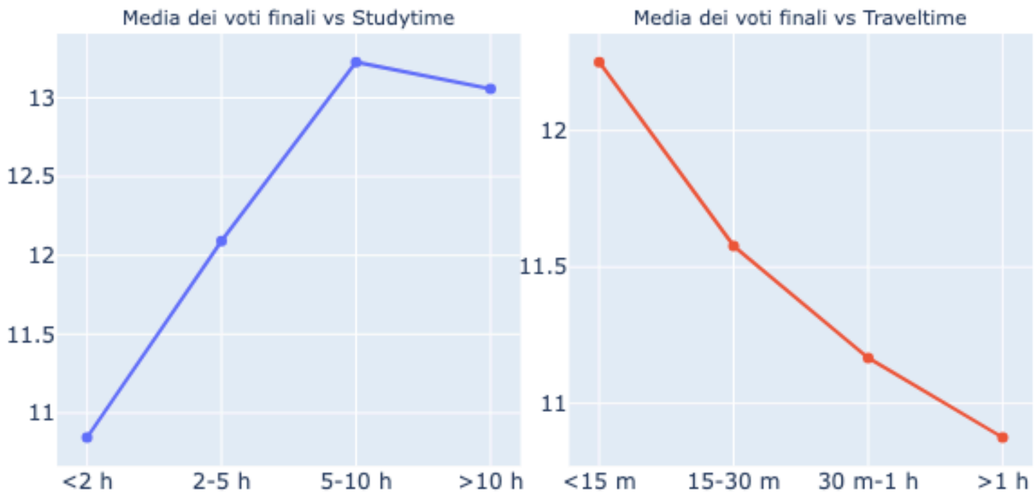
Ambiente Familiare

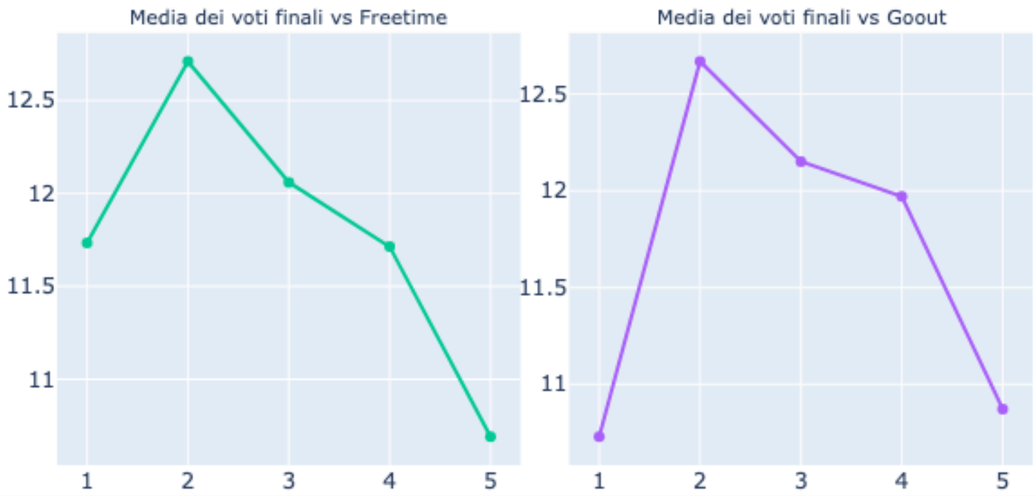
A seguire è stato cercato di far emergere come le performance accademiche degli studenti siano influenzate dal contesto familiare, con un focus specifico sulle categorie professionali e sul livello di istruzione dei genitori, notando come i figli di insegnanti abbiano rendimenti scolastici migliori, mentre i figli dei genitori che lavorano a casa, hanno rendimenti peggiori.





Gestione del tempo



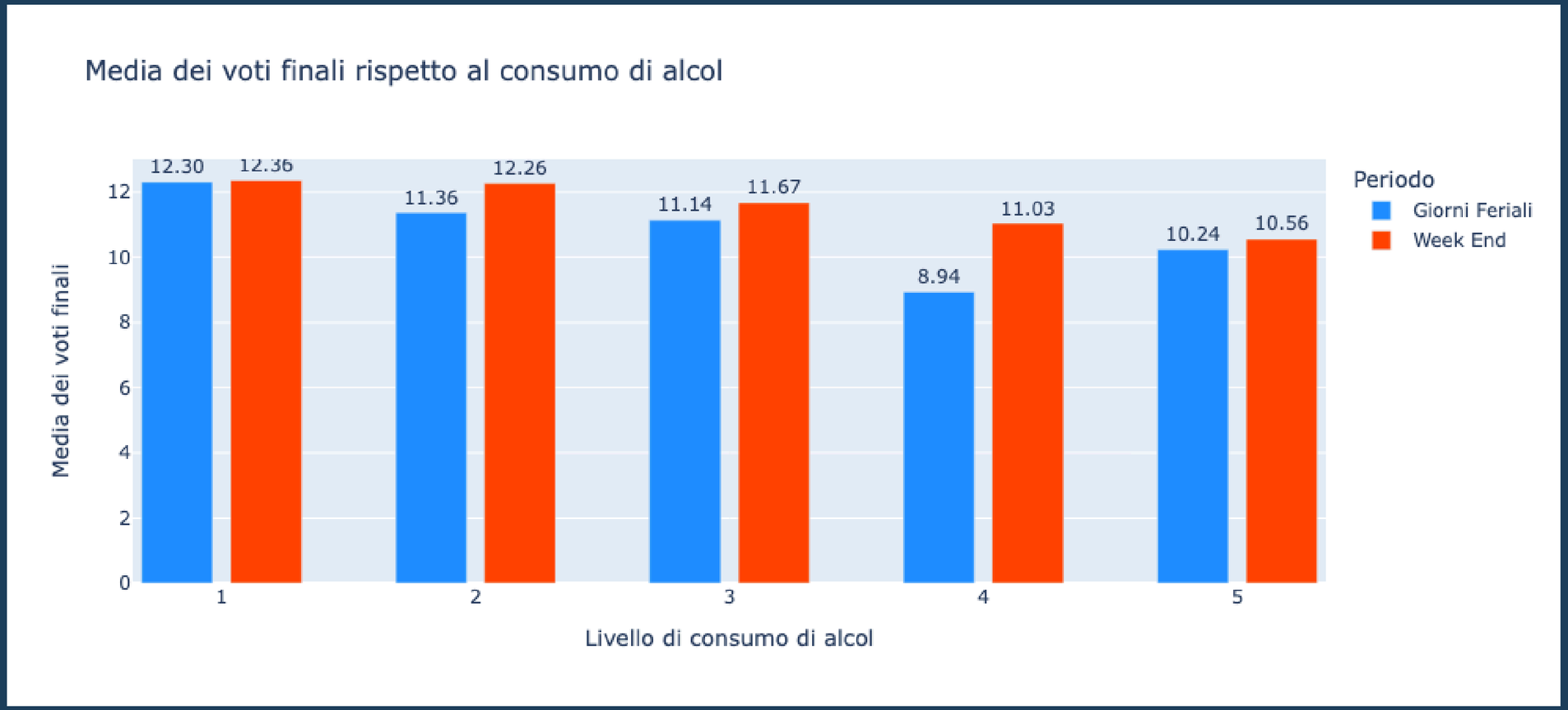


Questi grafici evidenziano:

- La correlazione positiva tra rendimento scolastico e quantità di studio, purché non si ecceda.
- La diminuzione dei voti all'aumentare della distanza tra casa e scuola.
- L'importanza del tempo libero e delle uscite con gli amici come validi alleati dell'apprendimento, se gestiti con equilibrio.

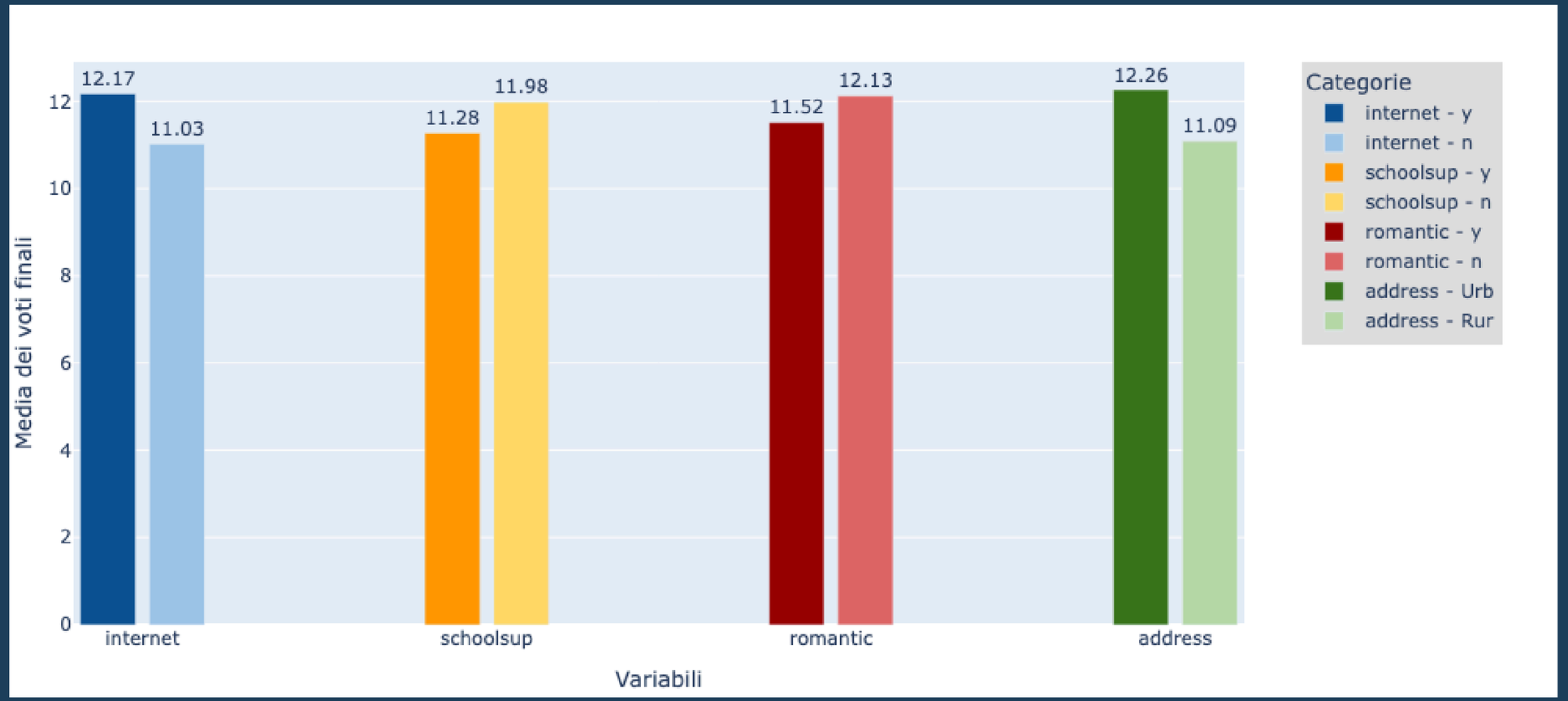
Consumo Alcolici

Questa analisi esplora la relazione tra il consumo di alcol nei giorni feriali e festivi e le medie dei voti finali.
L'analisi mostra che il consumo di alcol durante i giorni infrasettimanali è associato a voti finali più bassi rispetto al consumo nel weekend: questo indica un impatto negativo maggiore sulle prestazioni scolastiche nei giorni feriali.



Ulteriori Analisi

Questa sezione esamina come diverse variabili influenzano la media dei voti finali, facendo emergere che l'accesso a Internet e la residenza urbana sono positivamente correlati a voti finali più alti, mentre il supporto scolastico e le relazioni sentimentali possono influenzare negativamente le prestazioni accademiche



Analisi Predittiva

01

Panoramica delle Tecniche di ML

Definizione del problema

02

Algoritmi di ML

Algoritmi di ML selezionati per prevedere il voto.

03

Valutazione dei modelli

Valutazione e Paragone dei modelli di ML tramite metriche di regressione

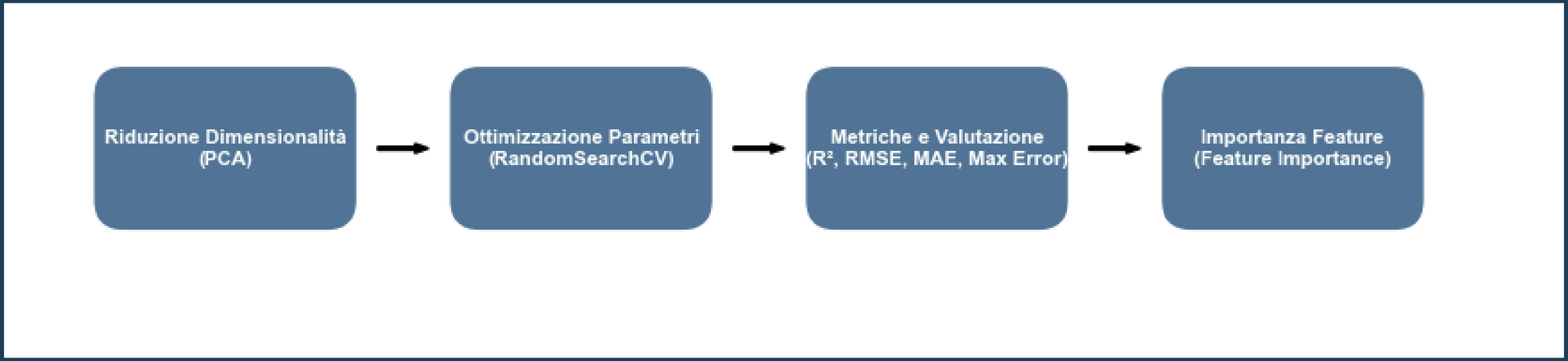
04

Importanza delle feature

Valutazione dell'importanza delle varie feature nei modelli di ML selezionati

Panoramica tecniche ML

In questa sezione, esamineremo i passaggi chiave per la predizione dei voti finali degli studenti utilizzando tecniche di machine learning.



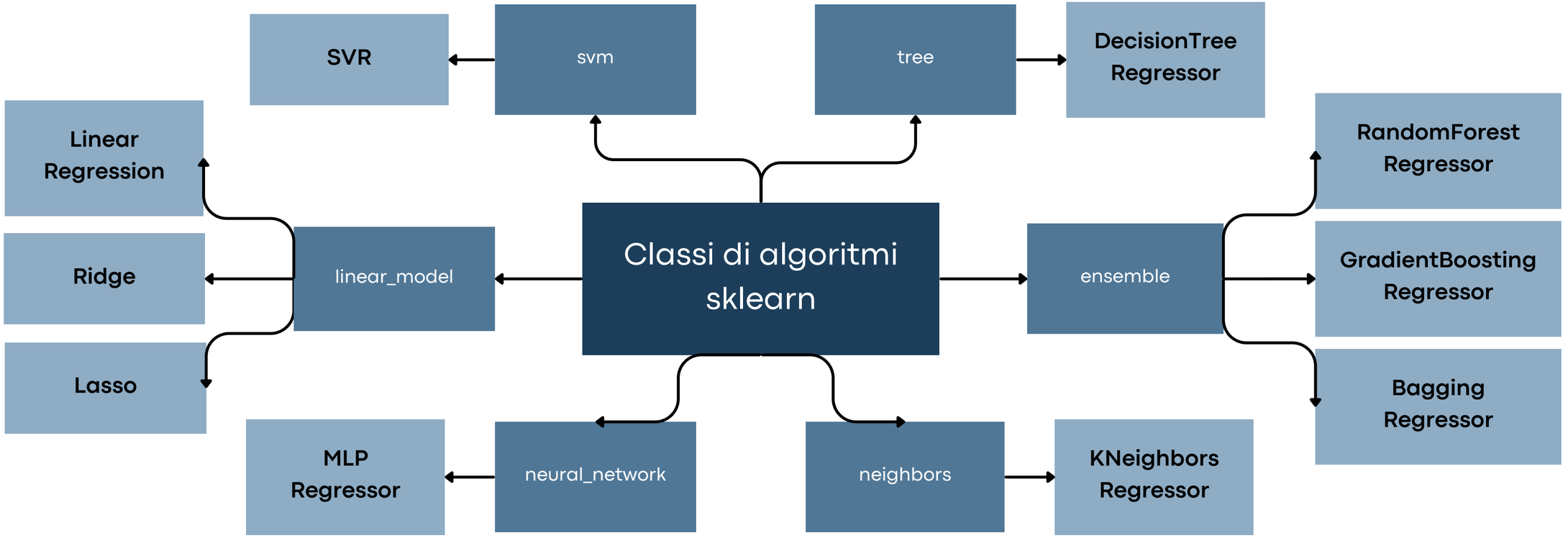
Algoritmi di ML Selezionati

Obiettivi

Per dimostrare l'efficacia di diversi algoritmi di machine learning nella previsione del voto finale, sono stati testati una gamma diversificata di modelli. Questo approccio didattico, ha permesso di:

- **Apprendere** come ogni algoritmo gestisce i dati e identifica le caratteristiche rilevanti
- **Valutare** le performance relative di ciascun modello, fornendo una comprensione pratica delle loro capacità e limitazioni.
- **Comprendere** le differenze fondamentali tra gli algoritmi, facilitando la scelta del modello più adatto per previsioni accurate

Questo diagramma a blocchi illustra la suddivisione degli algoritmi di ML in diverse classi principali, e i modelli specifici utilizzati per la previsione del voto finale. Ogni classe contiene uno o più algoritmi, che vengono esplorati per comprendere meglio le loro caratteristiche e performance.



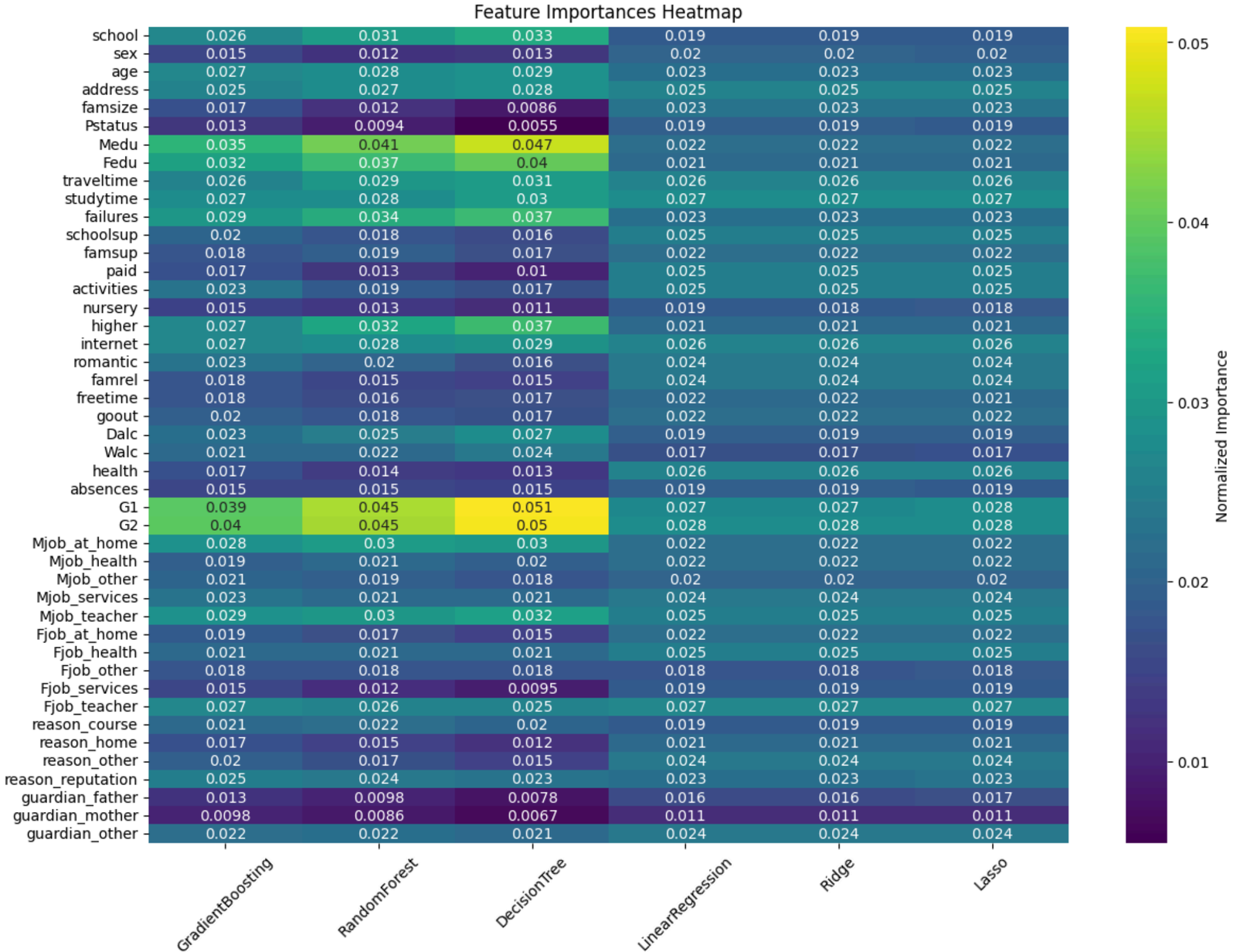
Valutazione

	R^2	RMSE	MAE	Max Err	Train Time (s)
SVR	0.8	1.4	0.92	9.98	2.18421
GradientBoosting	0.69	1.74	1.22	7.51	26.3019
Bagging	0.59	2.01	1.47	8.46	6.77059
RandomForest	0.6	1.98	1.45	8.55	6.24575
DecisionTree	0.45	2.32	1.76	10.26	0.157466
LinearRegression	0.8	1.41	0.96	9.99	0.0136209
Ridge	0.8	1.4	0.95	9.94	0.132025
Lasso	0.8	1.4	0.95	9.9	0.0934153
NeuralNetwork	0.71	1.67	1.19	9.44	20.1442
KNN	0.48	2.24	1.69	10.37	0.065825

L’analisi degli errori commessi dai vari modelli, ha fatto emergere una difficoltà degli algoritmi nel prevedere il voto 0

- **LinearRegression, Ridge, Lasso**(Modelli Lineari): Offrono prestazioni solide con valori di R^2 e RMSE tra i migliori. Sono anche molto veloci da addestrare.
- **SVR**: Ha prestazioni competitive, ma con tempi di addestramento più lunghi rispetto ai modelli lineari.
- **GradientBoosting e NeuralNetwork**: performance accettabili, ma il tempo di addestramento è molto più lungo. Potrebbero avere errori maggiori in alcune previsioni.
- **Bagging, RandomForest, Decision Tree e KNN**: Tendono ad avere prestazioni più basse con errori maggiori e un tempo di addestramento che varia. Questi modelli possono essere meno adatti per questo particolare problema di previsione del voto.

Importanza delle feature



Importanza delle Features

Considerazioni

Analisi per modelli

- **Linear Regression, Ridge e Lasso:** questi modelli mostrano che l'importanza complessiva delle features è più distribuita, riflettendo la loro natura lineare
- **Gradient Boosting, Random Forest, Decision Tree**: Questi modelli tendono ad assegnare importanze differenti alle features, riflettendo la loro capacità di catturare relazioni non lineari e interazioni tra feature.

Analisi per Features

- Le features come **G1** e **G2** hanno le importanze più alte in tutti i modelli, indicando che i voti precedenti sono i predittori più forti per il voto finale G3.
- Altre features che appaiono frequentemente con importanza significativa sono **Medu** e **Fedu**, suggerendo che il background educativo dei genitori è rilevante.
- **Pstatus**, **nursery** e **guardian** hanno importanze molto basse in tutti i modelli, suggerendo che non sono un fattore critico per predire i voti.