TREE BASED MODELS

Roberto Cerminara

Daniele Florio

Lorenzo Piattoli

DATASET

Student Performance Factors

Insights into Student Performance and Contributing Factors

Data Card | Code (203) | Discussion (1) | Suggestions (1)

About Dataset

Description

This dataset provides a comprehensive overview of various factors affecting student performance in exams. It includes information on study habits, attendance, parental involvement, and other aspects influencing academic success.

Usability ⓘ
10.00

License
CC0: Public Domain

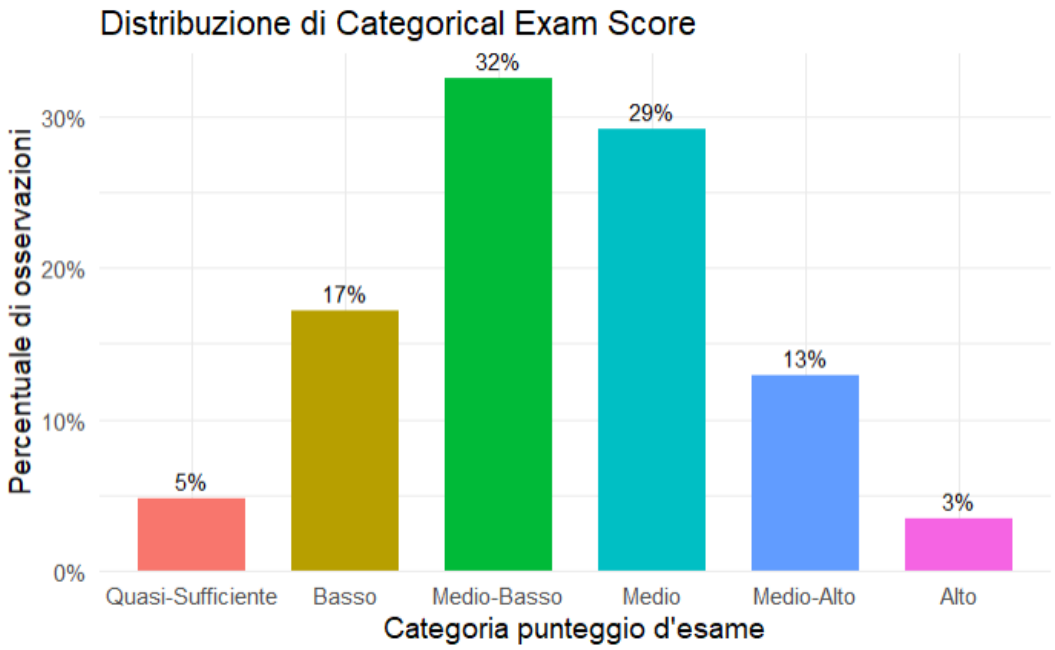
Expected update frequency

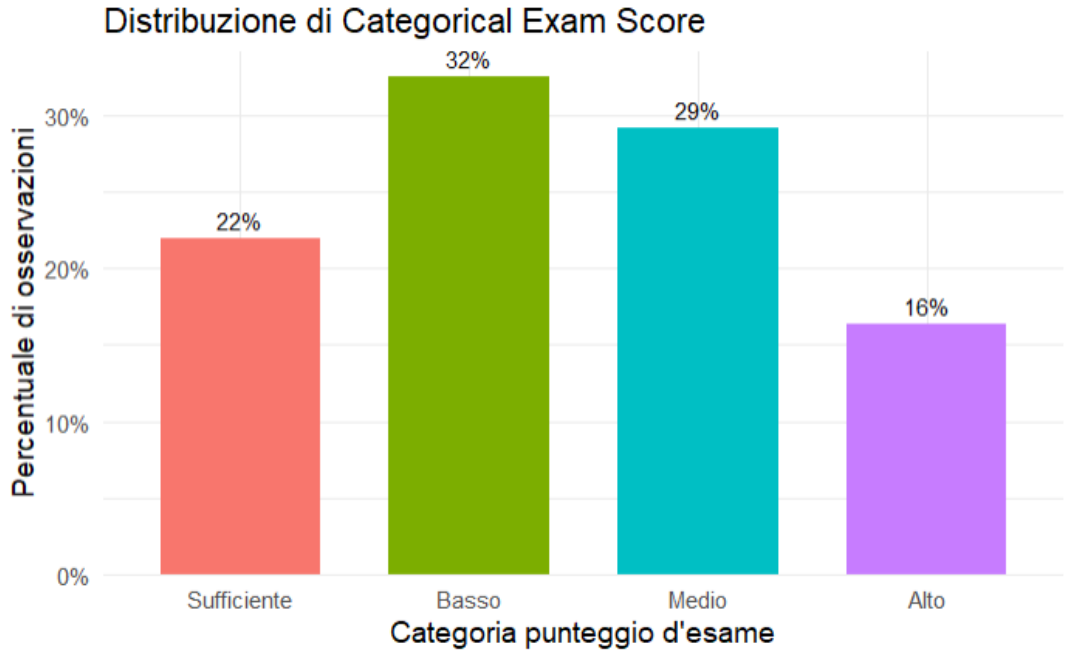
	Hours_Studied <int>	Attendance <int>	Parental_Involvement <fctr>	Access_to_Resources <fctr>	Extracurricular_Activities <fctr>	Sleep_Hours <int>	Previous_Scores <int>	Motivation_Level <fctr>
1	23	84	Low	High	No	7	73	Low
2	19	64	Low	Medium	No	8	59	Low
3	24	98	Medium	Medium	Yes	7	91	Medium
4	29	89	Low	Medium	Yes	8	98	Medium
5	19	92	Medium	Medium	Yes	6	65	Medium
6	19	88	Medium	Medium	Yes	8	89	Medium

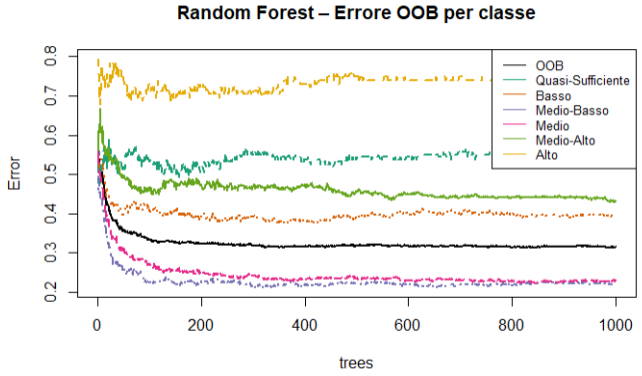
Link: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

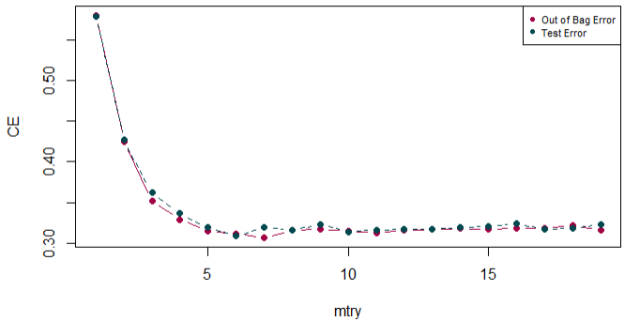
DATASET

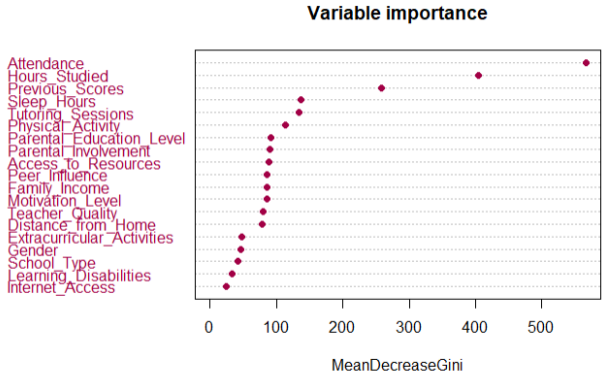
Sono stati generati due dataset con raggruppamenti diversi della variabile Exam Score (target):

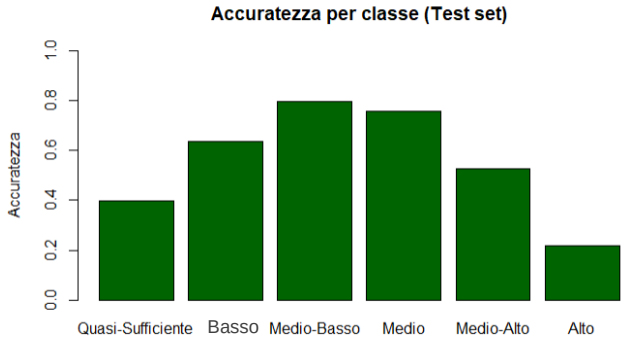










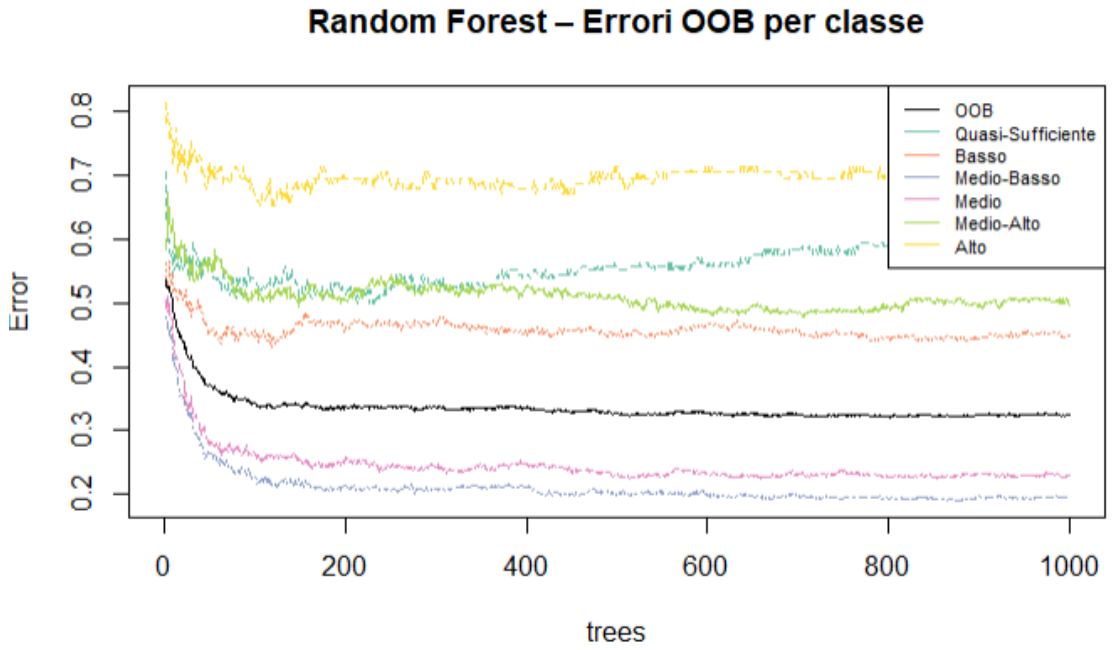


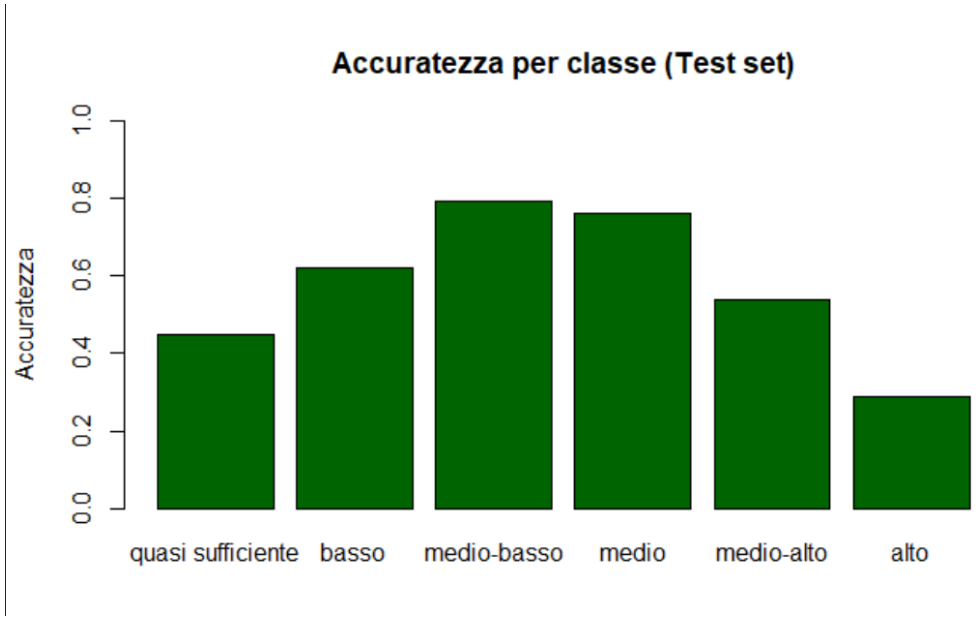
RANDOM FOREST

Vediamo i risultati principali del modello

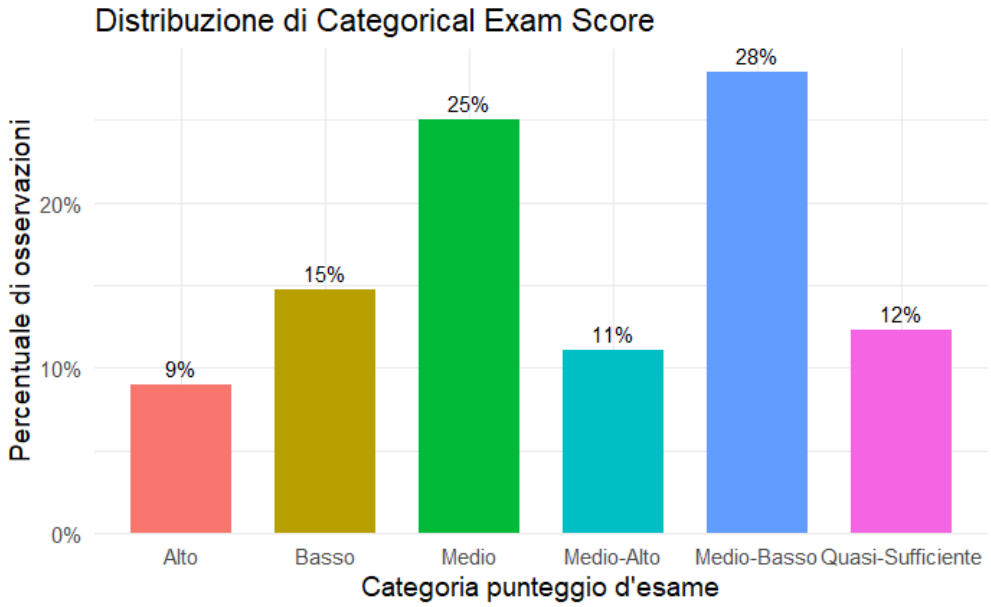
RANDOM FOREST - WEIGHTED

Comparazione del modello con classi pesate in base all'errore Out-of-bag e all'accuratezza sui dati di test

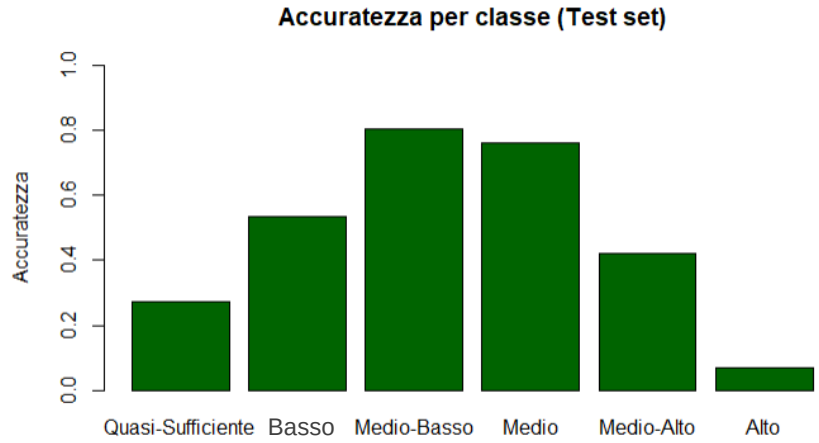




RANDOM FOREST SUL DATASET AUGMENTED



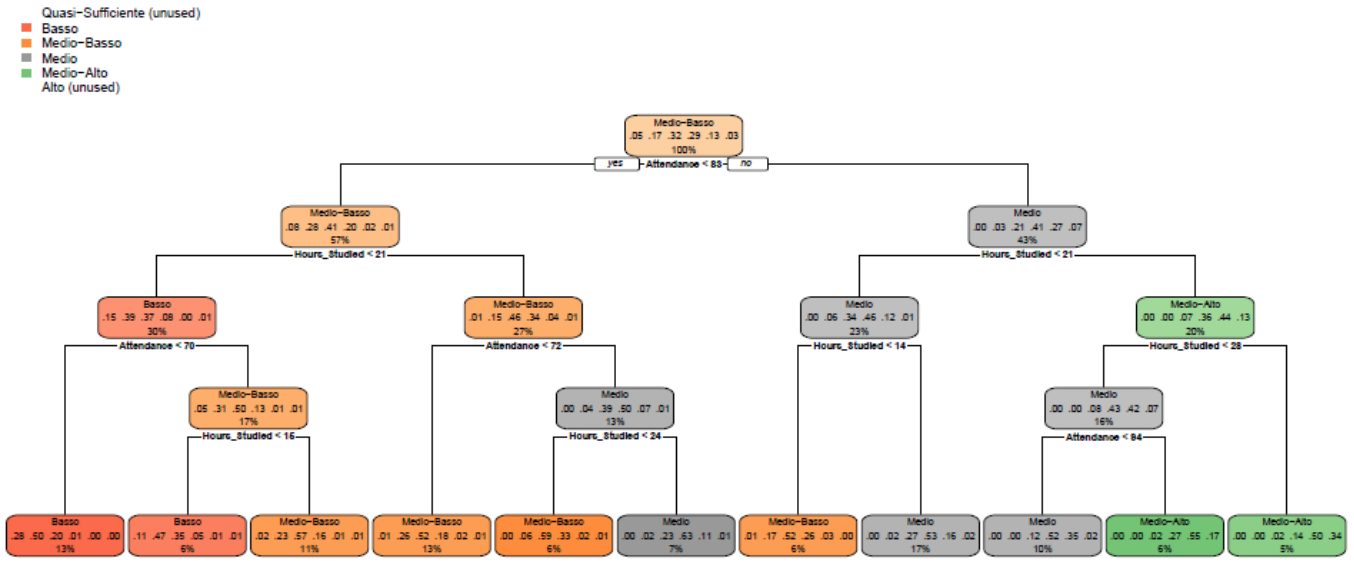
Predizione sui dati di test

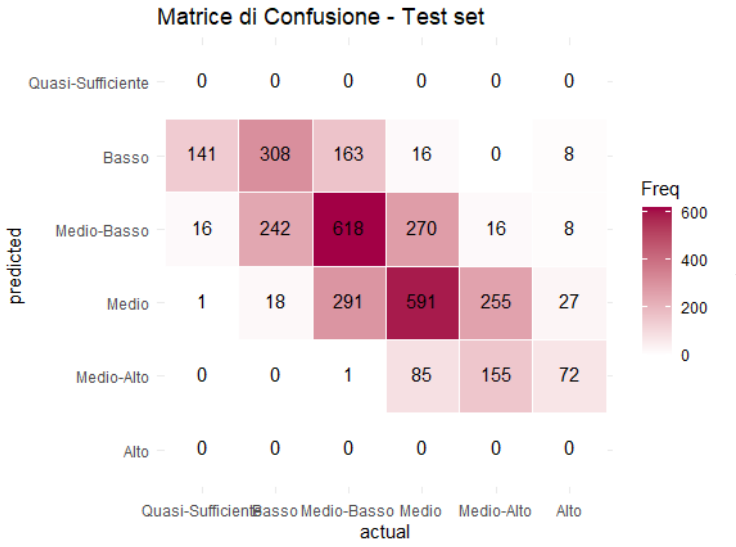


Distribuzione ottenuta dai dati aumentati

CART

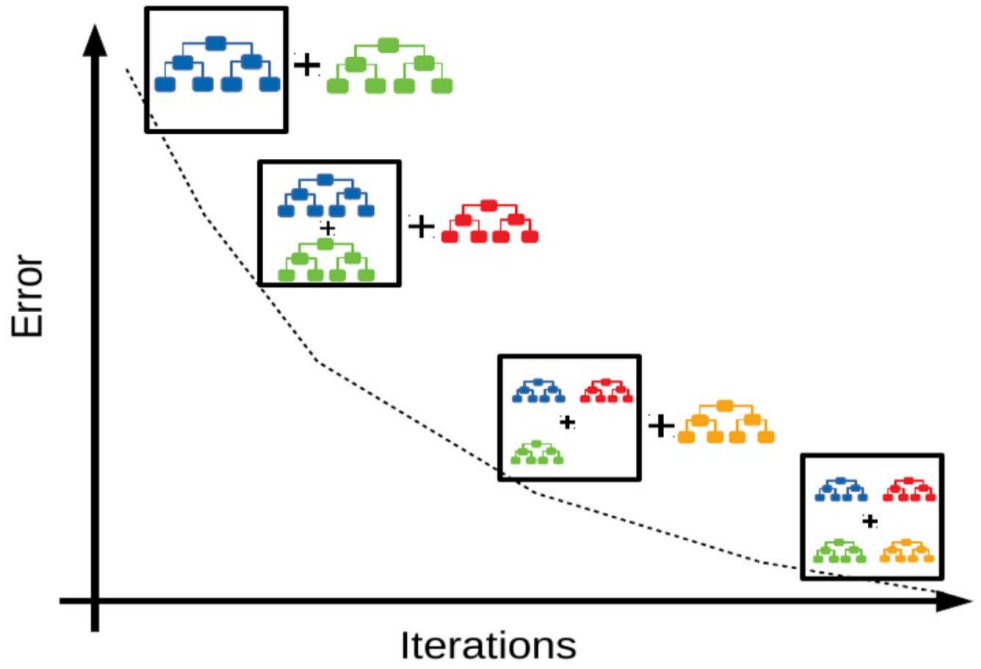
Risultato del CART



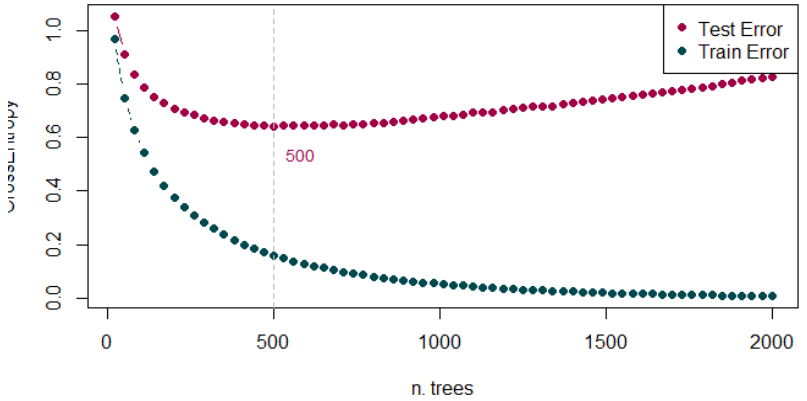


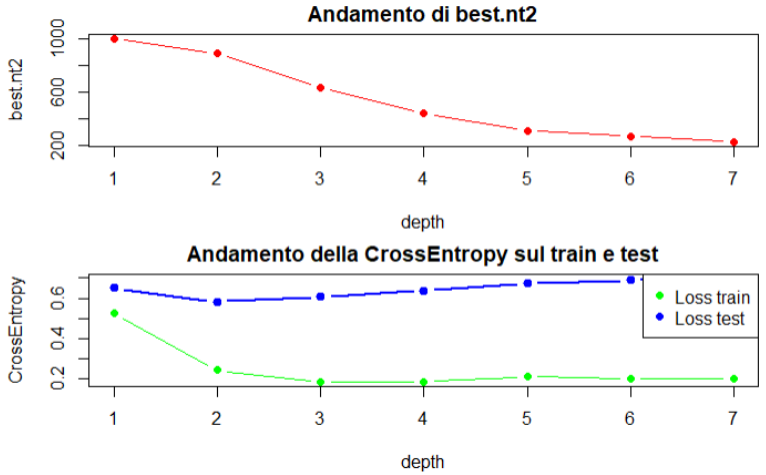
Si può osservare come il modello CART non classifica le categorie Alto e Quasi-Sufficiente

BOOSTING



Errore sul modello al variare del numero di alberi





Fine tuning dei parametri

CONCLUSIONI

Variable importance: Le variabili utilizzate con maggiore frequenza dagli algoritmi sono: le ore di studio settimanali, la percentuale di frequenza alle lezioni e i punteggi degli esami precedenti.

Performance dei modelli Random Forest: Senza pesi, il modello fatica a riconoscere le classi meno rappresentate, l'uso dei pesi riduce l'errore di queste ultime ma non porta a un miglioramento sostanziale. La classificazione in quattro fasce stabilizza il modello e ne aumenta l'accuratezza, mentre l'utilizzo dell'algoritmo SMOTE peggiora le prestazioni.

Altri modelli a confronto: Il modello CART è semplice ma meno efficace, con tendenza a sottostimare le classi estreme. Il Boosting, invece, si distingue per l'accuratezza (80% test) e la buona generalizzazione post-ottimizzazione.

Considerazioni: limitazioni legate alle dipendenze tra variabili, in particolare, modelli come CART soffrono quando si tratta di catturare relazioni lineari o prossime alla linearità, come evidenziato nel nostro caso di studio.