

Capitolo 1

Discussione e conclusioni

1.1 Discussione dei risultati

I risultati ottenuti ci mostrano come sia possibile predire l'abbandono di uno studente con affidabilità diverse in momenti diversi del primo anno di corso. Come prevedibile, è più facile predire l'abbandono di uno studente al passare del tempo, proprio per questo otteniamo risultati migliori nella predizione a fine del primo anno, risultati che peggiorano se consideriamo la fine del primo semestre e risultati ancora meno precisi al momento dell'iscrizione. Tra il modello predittivo a fine del primo anno e quello a fine primo semestre c'è una differenza di accuracy di circa l'11% con circa la stessa differenza anche nella recall dei dropout. La differenza in accuratezza tra il modello al momento dell'iscrizione ed il modello alla fine del primo semestre è di circa il 15% un valore non indifferente, mentre la recall varia di circa 10 punti percentuali. Un'altro risultato che possiamo notare è che non in tutti i casi un utilizzo di un maggior numero di dati (recuperando tuple riempiendo i NaN dell'ISEE come spiegato in precedenza) porta all'ottenimento di risultati migliori, poichè come possiamo vedere nel modello a fine del primo anno, con l'utilizzo del secondo bilanciamento i risultati peggiorano invece che migliorare. Possiamo quindi dire che il modello rispetta le attese, poichè è triviale capire che più dati relativi all'andamento scolastico dello studente (nel nostro caso il numero di CFU superati e il superamento o meno degli OFA) possa chiarire quello che è la possibile scelta di uno studente relativamente all'abbandono scolastico.

Infine utilizzando l'approccio dell'Ensamble Learning non abbiamo ottenuto un miglioramento considerevole delle performance. Specifichiamo però che questo è stato un primo approccio a questa metodologia e il miglioramento di esso lo lasciamo come sviluppo futuro.

1.2 Limiti del metodo

Sicuramente un limite del nostro modello è la perdita di dati. Dover bilanciare un dataset fortemente sbilanciato come quello utilizzato in questo lavoro porta ad una grande perdita di dati e di conseguenza informazioni che potevano migliorare/peggiore i risultati ottenuti. Un'altro importante limite del nostro modello è che si ottengono i risultati migliori alla fine del primo anno di corso, quando ormai, potrebbe essere troppo tardi per aiutare lo studente.

1.3 Lavori futuri

In questo lavoro abbiamo applicato un primo approccio ad una tecnica di Ensemble Learning, quale lo Stacking Algorithm, e lasciamo sicuramente come sviluppo futuro la rifinitura dell'approccio utilizzato che porterà sicuramente al miglioramento dei risultati ottenuti.

Per continuare questo lavoro sicuramente si potrebbe andare a rendere più specifico il modello rispetto ad alcune caratteristiche come il corso di studi o la scuola di appartenenza, per andare a valutare con più precisione il tasso e le motivazioni di abbandono legate ad ogni singolo corso di studio o ambito di studio. Anche l'utilizzo di nuove tecniche come l'utilizzo delle NB o del kNN come in [?] è un possibile sviluppo futuro. Un'altro possibile sviluppo futuro riguarda l'integrazione di alcuni dati che non vengono presi in considerazione in questo studio, come la situazione familiare o l'eventuale condizione di studente-lavoratore, che possono sicuramente incidere sul percorso di studi di un ragazzo.