IL POLITICO (Univ. Pavia, Italy)
2026, anno XCI, n. 1, pp. 205-216

PERFORMANCE DEGLI STUDENTI UNIVERSITARI: METODI DATA-DRIVEN PER IL SUPPORTO DECISIONALE

di Simone Gerzeli

Introduzione

Nel contesto delle trasformazioni contemporanee dei sistemi di istruzione superiore, la crescente disponibilità di dati amministrativi e accademici ha profondamente modificato le modalità di analisi dei processi formativi e delle traiettorie degli studenti. In tale prospettiva, i dati longitudinali assumono una rilevanza strategica, in quanto consentono di osservare nel tempo l'evoluzione dei percorsi accademici e di interpretare fenomeni complessi quali il successo formativo, il ritardo negli studi e l'abbandono universitario[1].

L'attenzione verso questi temi si inserisce in un più ampio dibattito sulla *governance* delle istituzioni educative, sempre più orientata a modelli decisionali basati su evidenze empiriche. In questo quadro, gli approcci *data-driven* si configurano come strumenti fondamentali per supportare processi decisionali a diversi livelli - amministrativo, gestionale e politico - contribuendo alla definizione di politiche universitarie più efficaci e mirate. Più recentemente, la letteratura sottolinea come tali approcci non si limitino alla previsione dei fenomeni, ma contribuiscano alla progettazione di interventi mirati e personalizzati per gli studenti a rischio di abbandono o ritardo del percorso di studi[2].

Dipartimento di Scienze Politiche e Sociali, Università di Pavia.

[1] O. Goren, L. Cohen, A. Rubinstein, *Early prediction of student dropout in higher education using machine learning models*, in B. Paasen, C. D. Epp, (eds.), "Proceedings of the 17th International Conference on Educational Data Mining" Atlanta Georgia USA, International Educational Data Mining Society, 2024.

[2] H.I. Patel, D. Patel, *From Data to Decisions: Enhancing Student Retention with*

Il presente lavoro si colloca all'intersezione tra analisi quantitativa e riflessione sulle politiche pubbliche dell'istruzione, proponendo l'applicazione di modelli di apprendimento automatico a dati longitudinali relativi agli studenti iscritti all'Università di Pavia nel periodo 2018-2022. I dati analizzati permettono di ricostruire le traiettorie educative lungo un arco temporale quinquennale e di individuare fattori di rischio e di successo, nonché di valutare l'efficacia di interventi volti alla riduzione del *dropout* e al miglioramento delle performance accademiche, confermando come i modelli predittivi basati su *machine learning* rappresentino oggi uno degli strumenti più promettenti per l'analisi delle carriere accademiche[3].

In questo contesto, l'utilizzo congiunto di modelli statistici tradizionali - quali il Modello Lineare Misto (LMM) e il Modello Lineare Generalizzato Misto (GLMM) - e di tecniche di apprendimento automatico - tra cui RANDOM FOREST, ADABOOST e XGBOOST - consente di mettere a confronto approcci metodologici differenti, evidenziandone potenzialità e limiti non solo in termini di capacità predittiva, ma anche rispetto alla loro interpretabilità e alla loro utilità nel supporto alle decisioni. La letteratura più recente evidenzia come tali modelli siano particolarmente efficaci nell'identificazione precoce degli studenti a rischio e nell'analisi di *dataset* complessi e multidimensionali[4].

La scelta dei modelli analitici non rappresenta, infatti, un aspetto meramente tecnico, ma si inserisce in una dimensione più ampia di costruzione della conoscenza e di orientamento delle politiche educative. In particolare, la sfida attuale consiste nel rendere i risultati dell'analisi non solo accurati, ma anche interpretabili e utilizzabili nei contesti decisionali, colmando il divario tra capacità predittiva e implementazione di politiche efficaci[5].

Predictive Analytics, in: T. SENJYU, C. SO-IN, A. JOSHI, (eds), "Smart Trends in Computing and Communications" Singapore, Springer, 2026.

[3] W. ZAMBRANO-ROMERO, C. RODRIGUEZ, J. PITA-VALENCIA, W.J. ZAMBRANO-ROMERO, J.M. MORAN-TUBAY, *Machine Learning in the Teaching Quality of University Teachers: Systematic Review of the Literature 2014-2024*, in "Information", n. 3, 2025.

[4] N. SGHIR, A. ADADI, M. LAHMER, *Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)*, in "Education and Information Technologies", n. 28, 2023.

[5] K. ALALAWI, R. ATHAUDA, R. CHIONG, *An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention*, in "International Journal of Artificial Intelligence in Education", n. 35, 2025.

L'analisi empirica si basa su un ampio *dataset* reale fornito dall'Università di Pavia, comprendente informazioni relative al superamento degli esami, alle aree disciplinari, al rendimento accademico e a variabili demografiche e sociali. Il campione include 231.740 osservazioni riferite a 53.726 studenti, permettendo una ricostruzione articolata dei comportamenti accademici nel tempo.

Nel loro insieme, i risultati del lavoro mirano a contribuire al dibattito sul ruolo degli strumenti quantitativi avanzati nella definizione delle politiche universitarie, mostrando come un approccio integrato tra analisi dei dati e comprensione dei contesti istituzionali possa favorire lo sviluppo di strategie più efficaci per il miglioramento dei sistemi educativi.

1. *Analisi della letteratura*

Le istituzioni educative dispongono oggi di un'ampia quantità di informazioni utili per analizzare le performance degli studenti e i risultati di apprendimento. In questo contesto, i dati longitudinali consentono di osservare l'evoluzione dei comportamenti accademici nel tempo, offrendo una base empirica solida per l'analisi dei percorsi formativi. In particolare, modelli a effetti misti e modelli di cambiamento (*change-point models*) sono stati utilizzati per valutare l'impatto della pandemia di COVID-19 sulle prestazioni accademiche degli studenti universitari[6].

Il Modello Lineare Generalizzato Misto (GLMM) e, più in generale, i modelli multilivello rappresentano strumenti consolidati per l'analisi di dati longitudinali, in quanto consentono di tenere conto della struttura gerarchica delle informazioni. Parallelamente, lo sviluppo di tecniche di apprendimento automatico basate su alberi decisionali ha ampliato le possibilità di analisi, rendendo possibile l'identificazione di *pattern* complessi nei dati[7].

[6] F. Naz, S. Gerzeli, E. Ballante, S. Figini, *Learning in lockdowns: a five-year analysis of covid-19's influence on university students' academic experiences*, in "Advances and applications in statistics", n. 91, 2024, pp 59-75.

[7] M. Fokkema, J. Edbrooke-Childs, M. Wolpert, *Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data*, in "Psychotherapy Research", n. 31, 2021, pp. 329-341.

Numerosi studi hanno applicato tali metodologie al contesto dell'istruzione superiore. In particolare, Behr et al.[8] mostrano come l'utilizzo di algoritmi di tipo RANDOM FOREST, basati su alberi di inferenza condizionale, consenta di prevedere in modo efficace l'abbandono universitario, sfruttando *dataset* articolati che includono variabili accademiche e socio-demografiche.

Analogamente, la letteratura evidenzia il crescente utilizzo di algoritmi di *boosting*, come XGBOOST, e di tecniche interpretative quali i valori SHAP, che permettono di comprendere il contributo delle variabili nei modelli predittivi[9].

Ulteriori contributi si concentrano sul sull'impatto delle strategie di apprendimento nel determinare il successo accademico. Neroni et al.[10] evidenziano come le modalità di studio influenzino significativamente le *performance* degli studenti, mentre Berens et al.[11] presentano metodi di apprendimento automatico per l'individuazione precoce di studenti a rischio di abbandono degli studi, utilizzando dati amministrativi delle carriere degli studenti.

Nel campo del *data mining* in ambito educativo, Francis e Babu[12] propongono un approccio basato su tecniche di classificazione e *clustering* per la previsione delle prestazioni accademiche degli studenti, mentre Gray e Perkins[13], analizzando il ruolo delle *learning analytics,* sviluppano un modello predittivo finalizzato all'individuazione precoce degli studenti in difficoltà.

[8] A. Behr, M. Giese, H. Teguim, K. Theune, *Early prediction of university dropouts–a random forest approach*, in "Jahrbucher fur Nationalokonomie und Statistik", n. 240, 2020, pp. 743-789.

[9] L. Hao, C. Xi, L. Xiaoxiao, *Factors influencing secondary school students' reading literacy: An analysis based on xgBoost and shap methods*, in "Frontiers in Psychology", n. 13, 2022.

[10] J. Neroni, C. Meijs, H. Gijselaers, P Kirschner, R Groot, L*earning strategies and academic performance in distance education. Learning and Individual Differences*, in "Learning and Individual Differences" n. 73, 2019, pp. 1-7. L

[11] J. Berens, K. Schneider, S. Görtz, S. Oster, J. Burghoff, *Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods,* in "Journal of Educational Data Mining", n. 11, 2019.

[12] F. Bindhia, S. Babu, *Predicting Academic Performance of Students Using a Hybrid Data Mining Approach*, in "Journal of Medical Systems", n. 43, 2019, pp. 1-15.

[13] C. Gray, D. Perkins, *Utilizing early engagement and machine learning to predict student outcomes*, in "Computers & Education", n. 131, 2019, pp. 22-32.

2. *Metodi e modelli*

I modelli di apprendimento automatico sono presentati per gestire dati longitudinali. Nel nostro lavoro, modelli parametrici basati sul Modello Lineare Misto (LMM) e Modello Lineare Generalizzato Misto (GLMM) sono confrontati con modelli di apprendimento automatico non parametrici come RANDOM FOREST, ADABOOST e XGBOOST.

2.1 *Modelli parametrici*

I LMM e i GLMM forniscono un quadro parametrico per spiegare una variabile dipendente Y sulla base di un insieme di covariate relative ai risultati accademici degli studenti. Nel *dataset* reale considerato, Y è il numero di crediti ottenuti dagli studenti durante l'anno accademico oppure, in un'applicazione secondaria, sull'evento di abbandono, Y è binaria e riflette il dropout.

Un modello misto può essere scritto come:

$$Y = X\beta + Zu + \varepsilon$$

dove:
Y rappresenta la variabile dipendente,
X è la matrice dei predittori,
β sono le stime dei coefficienti degli effetti fissi,
Z è la matrice degli effetti casuali,
u è il vettore dei coefficienti degli effetti casuali,
ε è il termine di errore residuo, che rappresenta la differenza tra il valore previsto e il valore osservato reale.

Naturalmente, se Zu è assente, il modello si riduce a un semplice modello lineare con soli effetti fissi.

2.2 *Modelli non parametrici*

In questo lavoro, per analizzare i dati disponibili, si sono adottati modelli non parametrici basati su alberi decisionali. RANDOM FOREST è una metodologia *ensemble* che può essere utilizzata per applicazioni predittive caratterizzate da Y quantitativa (cioè numero di crediti ottenuti da ciascuno studente) oppure Y qualitativa (cioè binaria per

l'evento di abbandono). Le statistiche di importanza delle variabili ottenute come output della RANDOM FOREST evidenziano i fattori più rilevanti rispetto al successo accademico.

ADABOOST è una tecnica di boosting, utile nell'analisi di dati longitudinali misti. Possiamo utilizzarla in compiti di classificazione come l'identificazione delle prestazioni accademiche degli studenti in termini di abbandono. L'equazione di ADABOOST può essere interpretata come segue:

$$F(x) = \Sigma_t \, \alpha_t \, h_t(x),$$

dove:
$F(x)$ rappresenta la funzione di previsione finale, che predice se uno studente abbandonerà o meno,
t è il numero totale di apprenditori deboli (singoli alberi decisionali) nell'ensemble,
α_t rappresenta il peso assegnato a ciascun apprenditore debole $h_t(x)$ in base alle sue prestazioni,
$h_t(x)$ è la previsione effettuata dal t-esimo apprenditore debole per l'input x, indicando la probabilità che uno studente abbandoni.

E' stato impiegato un altro modello di boosting, noto come XGBOOST, ovvero algoritmo di *extreme gradient boosting*. XGBOOST ha una varietà di utilizzi nell'analisi di dati longitudinali misti. In questo caso, XGBOOST può essere utile per determinare le variabili che possono influenzare i percorsi accademici degli studenti. L'individuazione di *pattern* nei dati longitudinali è resa più semplice con XGBOOST. XGBOOST si dimostra particolarmente efficace nell'individuazione di relazioni non lineari e pattern complessi nei dati longitudinali, il che aiuta a rivelare informazioni importanti.

L'equazione di XGBOOST può essere interpretata come segue:

$$Y_i = \Sigma_k \, f_k(x_i)$$

dove:
Y_i è il risultato previsto per l'i-esimo individuo,
k è il numero totale di apprenditori deboli (singoli alberi decisionali) nell'ensemble,
$f_k(x_i)$ è la previsione effettuata dal k-esimo apprenditore debole (albero) per l'input x_i.

3. *Risultati*

3.1 *Descrizione del dataset*

Il *dataset* è composto da 231.740 osservazioni relative a 53.726 soggetti distinti, con 70 variabili relative a diversi attributi degli studenti. I dati sono raccolti per un periodo temporale dal 2018 al 2022, complessivamente 10 semestri. Sono state considerate tre variabili target: crediti conseguiti (CFU), media dei voti e *dropout*.

Nei modelli statistici sviluppati sono state considerate come variabili indipendenti: anno, semestre, tipo di laurea e area disciplinare.

3.2 *Preprocessing dei dati*

Il *preprocessing* dei dati è una parte importante della nostra analisi. Poiché i dati erano in forma grezza, per renderli comprensibili in dettaglio abbiamo effettuato operazioni di pulizia e estrazione delle caratteristiche (*feature extraction*) per migliorare la capacità di generalizzazione del modello.

Il riepilogo dei dati è riportato di seguito. Per una migliore rappresentazione, i periodi accademici - che erano nella forma di due semestri nell'arco di cinque anni - sono stati suddivisi in dieci semestri accademici.

Per semplificare l'analisi, i diversi ambiti di studio sono stati combinati in cinque principali aree: Ingegneria, Umanistica, Giuridico-Economico-Politica, Medica, Scientifica, mentre i diversi Corsi di laurea sono stati combinati in quattro categorie principali: laurea triennale, laurea magistrale, laurea ciclo unico di 5 anni, laurea ciclo unico di 6 anni. La distribuzione degli studenti per tipologia di laurea e per area disciplinare è riportata nelle Tabelle 1 e 2.

Tabella 1 - *Frequenza degli studenti per tipo di laurea.*

Tipo di laurea	*Numero di studenti*
Triennale	13.776
Magistrale	4.384
Ciclo unico 5 anni	3.174
Ciclo unico 6 anni	1.840

Tabella 2 - *Frequenza degli studenti per area disciplinare.*

Area disciplinare	*Numero di studenti*
Giuridico-Economico-Politica	6.416
Medica	3.496
Scientifica	4.376
Ingegneria	876
Umanistica	2.591

La distribuzione degli esami sostenuti dagli studenti è stata rappresentata in forma binaria, dove: valore 0 significa che nessun esame è stato sostenuto, valore 1 significa che sono stati sostenuti esami durante i dieci semestri accademici considerati.

Gli abbandoni degli studenti (*dropout*) nel corso dei dieci semestri accademici sono stati classificati in categorie binarie (0 e 1).

Gli studenti che hanno partecipato a programmi all'estero (*Erasmus*) non sono stati inclusi, al fine di evitare errori nell'assegnazione degli esami e dei crediti tra i semestri.

L'analisi del *dropout* fornisce informazioni sui fattori che influenzano il percorso educativo di uno studente. La valutazione dei crediti conseguiti dagli studenti (CFU) offre una visione completa del progresso del percorso accademico.

La media dei voti è un'altra dimensione cruciale di *performance*, utilizzata per analizzare le tendenze e le variazioni nelle prestazioni accademiche degli studenti. L'analisi di queste variazioni può aiutare a identificare i *pattern* e i fattori che contribuiscono al rendimento accademico complessivo di uno studente.

3.3 *Confronto tra i diversi modelli*

Sono state condotte analisi empiriche comparative per esaminare i *pattern* dei *dropout*, dei crediti (CFU) e della media dei voti degli studenti dal 2018 al 2022 (10 semestri), implementando diversi modelli. L'analisi ha riguardato le relazioni tra le variabili *target* (*dropout*, CFU e media dei voti) e le principali variabili esplicative quali il numero di semestri, il tipo di corso di laurea e l'area disciplinare.

Il *dataset* è stato suddiviso in tre parti: set di addestramento, set

di test e set di validazione. Il settanta percento dei dati è stato utilizzato per il set di addestramento, il venti percento per il set di test e il restante dieci percento per il set di validazione.

Per quanto riguarda i modelli parametrici, il Linear Mixed Model (LMM) è stato applicato per la previsione dei CFU e della media dei voti, trattandosi di variabili continue, mentre il Generalized Linear Mixed Model (GLMM) è stato utilizzato per la previsione del *dropout*, in quanto variabile binaria.

I modelli sono stati valutati mediante metriche standard di *performance*, quali *Accuracy, Root Mean Square Error* (RMSE) e *Area Under the Curve* (AUC), al fine di garantire una valutazione comparativa robusta.

I risultati mostrano che il modello LMM presenta buone capacità predittive, con valori di RMSE pari a 0,78 sul training set e 0,80 sul test set per i CFU, e 0,70 e 0,73 rispettivamente per la media dei voti, senza variazioni significative nella fase di validazione.

Analogamente, il modello GLMM ha evidenziato una buona capacità di classificazione dei *dropout*, con un'accuratezza pari a 0,87 sul *training set* e 0,82 sul *test set*, e un valore di AUC pari a 0,83.

Per quanto riguarda i modelli non parametrici, i risultati evidenziano prestazioni complessivamente superiori, in particolare per quanto concerne l'identificazione precoce degli studenti a rischio.

Il modello RANDOM FOREST ha ottenuto i migliori risultati nella previsione dei *dropout*, con un'accuratezza pari a 0,85 sul training set e 0,83 sul *test set*, oltre a un valore di AUC pari a 0,92, indicando un'elevata capacità discriminante.

Il modello ADABOOST, pur mostrando buone prestazioni sul training set (0,83), evidenzia una minore capacità di generalizzazione, con una riduzione dell'accuratezza sul test set (0,78).

Per quanto riguarda i modelli di regressione non parametrici, XGBOOST ha mostrato prestazioni significativamente migliori rispetto ai modelli parametrici, con valori di RMSE molto contenuti (0,06 sul training set e 0,08 sul test set per i CFU), confermando la capacità di tale approccio di catturare relazioni non lineari e pattern complessi nei dati longitudinali. I risultati dei modelli sono riassunti nella Tabella 3.

Tabella 3 - *Risultati dei modelli parametrici e non parametrici.*

	CFU		
Modello	*RMSE Train*	*RMSE Test*	*Validation*
LMM	0,78	0,8	0,82
XGBOOST	0,06	0,08	0,12
	Media voti		
Modello	*RMSE Train*	*RMSE Test*	*Validation*
LMM	0,7	0,73	0,73
XGBOOST	0,07	0,09	0,13
	Dropout		
Modello	*Accuracy Train*	*Accuracy Test*	*Validation/AUC*
RANDOM FOREST	0,85	0,83	0,81 / 0,92
GLMM	0,87	0,82	0,79 / 0,83
ADABOOST	0,83	0,78	0,77 / 0,87

L'analisi evidenzia che i modelli parametrici e non parametrici presentano caratteristiche complementari. I modelli parametrici (LMM e GLMM) offrono una maggiore interpretabilità dei risultati, risultando particolarmente utili per l'analisi delle relazioni tra variabili e per la comprensione dei meccanismi sottostanti i fenomeni osservati. Al contrario, i modelli non parametrici, pur risultando meno interpretabili, mostrano una capacità predittiva superiore, soprattutto nella gestione di relazioni complesse e non lineari.

In particolare, RANDOM FOREST si distingue per l'elevata capacità di classificazione dei *dropout*, mentre XGBOOST risulta particolarmente efficace nell'ambito di problemi di regressione, confermando quanto evidenziato dalla letteratura recente in ambito di *machine learning* applicato ai dati educativi.

Questi risultati suggeriscono che la scelta del modello non può essere guidata esclusivamente da criteri di *performance* predittiva, ma deve tenere conto anche della finalità dell'analisi, distinguendo tra obiettivi di previsione e obiettivi di interpretazione. La scelta del modello deve essere guidata sia dalle caratteristiche del *dataset* sia dagli obiettivi conoscitivi e interpretativi della ricerca.

Osservazioni conclusive

In questo lavoro si è analizzata l'applicazione di modelli parametrici e non parametrici per lo studio dei dati longitudinali relativi ai percorsi accademici degli studenti universitari, evidenziandone potenzialità e limiti in una prospettiva comparativa.

I risultati ottenuti mostrano chiaramente che i modelli di apprendimento automatico, in particolare RANDOM FOREST e XGBOOST, offrano prestazioni superiori rispetto ai modelli tradizionali in termini di capacità predittiva. In particolare, RANDOM FOREST si conferma il modello più efficace per la previsione del *dropout*, grazie all'elevata accuratezza e al valore di AUC significativamente superiore rispetto agli altri approcci. Analogamente, XGBOOST risulta il modello più performante per la previsione dei CFU e della media dei voti, evidenziando la sua capacità di modellare relazioni complesse nei dati longitudinali.

Tuttavia, i risultati evidenziano anche l'importanza di considerare il *trade-off* tra capacità predittiva e interpretabilità. I modelli parametrici, pur meno performanti, mantengono un ruolo centrale nelle analisi orientate alla comprensione dei fenomeni, risultando particolarmente utili in contesti decisionali in cui la trasparenza e la capacità esplicativa dei risultati costituiscono requisiti fondamentali.

Dal punto di vista delle politiche universitarie, l'integrazione tra approcci *data-driven* e conoscenza del contesto istituzionale emerge come un elemento chiave per la definizione di strategie efficaci. L'identificazione precoce degli studenti a rischio, supportata da modelli predittivi accurati, può infatti consentire l'implementazione di interventi mirati di supporto, contribuendo alla riduzione dei tassi di abbandono e al miglioramento delle performance accademiche.

Le variabili più rilevanti individuate - area disciplinare, tipologia di corso di laurea e progressione temporale (numero di semestri) - confermano il carattere multidimensionale dei percorsi accademici e sottolineano la necessità di politiche differenziate, capaci di adattarsi alle specificità dei diversi gruppi di studenti.

In prospettiva futura, ulteriori sviluppi della ricerca potrebbero riguardare l'integrazione tra modelli ensemble e modelli gerarchici, nonché l'utilizzo di tecniche di *explainable AI*, al fine di migliorare la trasparenza dei modelli predittivi e favorire la loro applicazione per lo sviluppo di politiche universitarie efficaci, trasparenti e orientate al miglioramento dei percorsi formativi.

Abstract - This article analyses longitudinal data concerning university students through statistical and machine learning models, with the aim of evaluating academic performance, dropout risk and educational pathways within higher education systems. The study is based on a real dataset provided by the University of Pavia, including 231,740 observations related to 53,726 students over the period 2018–2022.

The research compares parametric approaches, such as Linear Mixed Models (LMM) and Generalized Linear Mixed Models (GLMM), with non-parametric ensemble methods, including RANDOM FOREST, ADABOOST and XGBOOST. The findings show that machine learning models, particularly RANDOM FOREST and XGBOOST, provide higher predictive performance in identifying students at risk of dropout and in forecasting academic outcomes. At the same time, parametric models maintain greater interpretability and explanatory capacity, which remain essential for institutional decision-making processes.

The article highlights the relevance of data-driven approaches for supporting university governance and educational policies, emphasizing the need to balance predictive accuracy, interpretability and policy applicability. The results suggest that the integration between advanced quantitative methods and institutional knowledge can contribute to the development of more effective strategies aimed at improving student retention and academic success.