UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO JONICO IN SISTEMI GIURIDICI ED ECONOMICI DEL MEDITERRANEO SOCIETÀ, AMBIENTE, CULTURE

IONIAN DEPARTMENT OF LAW, ECONOMICS AND ENVIRONMENT

ANNALI 2017 DEL DIPARTIMENTO JONICO

ANNO V

ESTRATTO

ANTONELLA SERRA
PAOLA PERCHINUNNO • MASSIMO BILANCIA

Previsione dell'abbandono degli studi universitari mediante algoritmi di *machine learning*: un caso di studio su dati dell'Università degli Studi di Bari

http://edizionidjsge.uniba.it/ • ISBN - 978-88-9095-699-7

EDIZIONI DJSGE

Contatti:

Prof. Nicola Triggiani
Dipartimento Jonico in Sistemi Giuridici ed Economici del Mediterraneo: società, ambiente, culture
Via Duomo, 259 - 74123 Taranto, Italy
E-mail: annali.dipartimentojonico@uniba.it
Telefono: + 39 099 372382
Fax: + 39 099 7340595
http://edizionidjsge.uniba.it/

A. Serra, P. Perchinunno, M. Bilancia

PREVISIONE DELL'ABBANDONO DEGLI STUDI UNIVERSITARI MEDIANTE ALGORITMI DI MACHINE LEARNING: UN CASO DI STUDIO SU DATI DELL'UNIVERSITÀ DEGLI STUDI DI BARI[*] [**]

ABSTRACT	
L'obiettivo di questo lavoro è quello di prevedere, in modo del tutto a-teoretico, gli studenti che sono a rischio di abbandono degli studi universitari. I dati utilizzati in questo lavoro riguardano l'Università degli Studi di Bari nel periodo 2013-2016, e sono stati estratti dall'Osservatorio Studenti-Didattica del Miur Cineca. L'analisi dei dati è basata esclusivamente sulle informazioni di profilazione che sono disponibili per ciascuno studente nel sistema informativo di Ateneo. La previsione degli abbandoni è stata effettuata utilizzando un insieme di tecniche di Machine Learning, note come algoritmi di classificazione supervisionata.	The aim of this paper is to predict, on a purely algorithmic basis, students who are at risk of dropping out of university. Data used in this study originated from the University of Bari Aldo Moro, during 2013-16, and were provided by the Osservatorio Studenti-Didattica of Miur-Cineca. Data analysis is based solely on the information set available, for each student, inside the university information system. Predictions of individual dropouts have been carried out by means of a subset of suitable Machine Learning techniques, known as supervised classification algorithms.
Abbandono degli studi universitari – Machine Learning – Previsione	**Students' dropout at university – Machine Learning – Prediction**

[*] Saggio sottoposto a referaggio secondo il sistema del doppio cieco.

[**] La presente ricerca è stata svolta all'interno delle attività del programma di ricerca n.13/08 del Dipartimento Jonico dell'Università degli Studi di Bari Aldo Moro, dal titolo: "Analisi di Business Intelligence finalizzate ad intercettare gli abbandoni da parte degli studenti universitari iscritti al Dipartimento Jonico", inserito all'interno dell'accordo di programma con il Comune di Taranto 2011/2013 per il consolidamento del Polo Universitario Jonico. Desideriamo ringraziare il Magnifico Rettore dell'Università degli Studi di Bari per avere autorizzato la consultazione, in forma anonima e a soli scopi di ricerca scientifica, dei dati relativi alla popolazione studentesca UniBA dell'Osservatorio Studenti-Didattica del MIUR-Cineca. Desideriamo infine ringraziare il dott. Massimo Iaquinta, responsabile dell'U.O. Statistiche di Ateneo, Settore Servizi Istituzionali, per il supporto offerto al ritrovamento dei dati. Gli autori hanno collaborato in parti uguali alla stesura del presente saggio.

1. Negli ultimi anni il sistema universitario italiano è stato sottoposto a procedure trasparenti di valutazione e monitoraggio, grazie ad opportune misure e norme varate dai diversi governi che si sono succeduti. L'Anvur (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca), attraverso il Rapporto sullo Stato del Sistema Universitario e della Ricerca[1], raccoglie e analizza le informazioni necessarie al monitoraggio periodico delle attività didattiche e scientifiche del sistema universitario e della ricerca. Tra le principali criticità emerse da tale rapporto si evidenzia come il fenomeno degli abbandoni degli studi universitari rappresenti uno dei problemi più rilevanti da affrontare, anche in termini di confronto con il sistema universitario europeo. Anche il Cnvsu[2] (Comitato Nazionale per la Valutazione del Sistema Universitario, istituito dal Ministero dell'Istruzione, dell'Università e della Ricerca) ha evidenziato che il tasso di abbandono studentesco deve essere considerato come un indicatore critico ai fini della valutazione degli Atenei e del Sistema Universitario nel suo complesso.

Ai fini della valutazione del sistema universitario appare estremamente significativo analizzare soprattutto la percentuale di abbandoni tra il primo ed il secondo anno di corso. La letteratura di riferimento sul tema e i risultati emersi da numerose indagini specifiche sugli abbandoni del sistema universitario evidenziano, infatti, che si tratta di un momento cruciale nel percorso degli studenti, con riflessi importanti sull'esito della propria carriera di studio. Gli ultimi dati disponibili rilevati dall'Anvur nell'anno accademico 2014/15 (disponibili sul Rapporto 2016) presentano una riduzione del tasso di abbandono negli ultimi anni e assumono dimensioni differenti se si disaggrega il dato per tipologia di corso di studio, distinguendo il tasso di abbandono delle lauree triennali da quello delle specialistiche (magistrali). In particolare, per quanto riguarda le lauree triennali tra il primo e il secondo anno di corso il tasso di abbandono è pari al 13.7%; tale percentuale dopo il terzo anno di corso sale al 24.7% per arrivare all'undicesimo anno al 38.7%. La situazione dei corsi biennali di secondo livello appare decisamente migliore in quanto il tasso di abbandono tra il primo e secondo anno risulta pari all'8%, salendo al 20.4% dopo l'undicesimo anno.

Le caratteristiche in ingresso degli studenti (genere, età, titolo di studio, voto del diploma) condizionano l'esito del percorso accademico, causando il fenomeno della dispersione universitaria (inattività, ritardo, abbandono) legato, spesso, alle difficoltà da parte dello studente, a confrontarsi con il mondo accademico e con la tipologia di corso in cui ha scelto di immatricolarsi. Particolarmente rilevante appare, infatti, l'influenza del percorso di studi da cui provengono gli studenti: nei corsi di laurea triennali il tasso di abbandono tra primo e secondo anno è pari all'8% degli studenti

[1] Anvur, 2016, http://www.anvur.org/attachments/article/1045/ANVUR_Rapporto_INTEGRALE_~.pdf.
[2] Cnvsu, 2011, http://www.cnvsu.it/_library/downloadfile.asp?id=11778.

provenienti dal liceo, contro il 20% circa degli studenti provenienti da un istituto tecnico e il 28% di quelli provenienti da un istituto professionale. Questo tipo di approccio all'analisi del problema è identico a quello di numerosi studi scientifici presenti in letteratura, dedicati al tentativo di determinare esplicitamente le determinanti e i fattori che portano all'insuccesso degli studi universitari[3]. Da tempo, molti Atenei italiani stanno affrontando questi temi, nel tentativo di contrastare i fenomeni di rallentamento e ridefinizione dei percorsi, alla ricerca delle principali cause che incidono sull'abbandono, sul ritardo o sul cambio di percorso negli studi universitari e di quali correttivi potrebbero dimostrarsi efficaci in tal senso.

Il punto di vista che adottiamo in questo lavoro è alquanto differente. Invece di teorizzare sulle possibili determinanti dell'abbandono, il nostro obiettivo è quello di prevedere, in modo del tutto a-teoretico, gli studenti che sono a rischio di abbandono, prima che tale abbandono sia stato effettivamente posto in essere e formalizzato[4,5]. La previsione deve essere basata esclusivamente sulle informazioni di profilazione che sono disponibili per ciascuno studente nel sistema informativo di Ateneo. In questo modo, l'obiettivo è quello di fornire un sistema di supporto alle decisioni che permetta di individuare precocemente gli studenti più a rischio, al fine di poter mettere in atto misure correttive che possano contribuire a ridurre il fenomeno oggetto di studio. L'approccio previsivo costituisce, pertanto, una sorta di 'prevenzione primaria' del fenomeno degli abbandoni dell'istruzione superiore.

Questo modo specifico di approcciarsi al problema degli abbandoni è parte integrante di un più vasto campo di ricerca emerso negli ultimi anni, che prende il nome di '*Educational Data Mining*' (EDM). Il processo di Data Mining, noto anche come '*Knowledge Discovery in Databases*' (KDD), consiste nel ritrovamento automatico, attraverso opportuni algoritmi, di informazioni nuove e potenzialmente utili che sono nascoste all'interno di grandi quantità di dati. Queste tecniche hanno trovato applicazione in un gran numero di campi, tra i quali possiamo citare il marketing, la bioinformatica e la previsione degli eventi di natura terroristica. Negli ultimi anni, è invece cresciuto esponenzialmente il loro utilizzo all'interno della ricerca scientifica sperimentale in campo educativo[6]. L'EDM è appunto quell'area di indagine scientifica incentrata sullo sviluppo di metodi ad hoc utilizzabili per scoprire regolarità e nuove informazioni all'interno di basi di dati provenienti da contesti legati all'istruzione, con lo scopo di poter comprendere meglio i singoli studenti, nonché gli ambienti all'interno dei quali tale istruzione viene erogata, e la loro relazione rispetto alle performance e agli obiettivi attesi[7]. In letteratura sono disponibili una serie di interessanti lavori di rassegna, che possono costituire un buon

[3] Bennett, 2003, 123-141.
[4] Di Pietro, 2004, 187-191.
[5] Belloc *et al.*, 2009, 127-138.
[6] Mohamad, Tasir, 2013, 320-324.
[7] Baker, 2017, *to appear*.

punto di partenza per gli studiosi specificatamente interessati ad approfondire le applicazioni in questo campo specifico[8,9,10].

Mentre il Data Mining è un processo che include anche l'estrazione, il pre-processamento e la presentazione dei dati, lo studio degli algoritmi che estraggono nuova conoscenza dalle basi di dati è delegato a quel campo generalmente indicato con il termine '*Machine Learning*' (ML, apprendimento automatico)[11]. Una classe importante di algoritmi di ML, che utilizzeremo in questo lavoro, data la struttura del problema che intendiamo studiare, è costituita dagli *algoritmi di classificazione supervisionata*. Essi si basano sulla disponibilità di un *training set* ad informazione completa, nel quale per ogni esempio (istanza) del problema oggetto di studio è disponibile tanto l'etichetta di classificazione (di solito si tratta di un'etichetta binaria 0/1), quanto un insieme di valori di s variabili di input qualitative/quantitative. Sulla base di questo insieme, l'algoritmo apprende una relazione empirica tra lo spazio delle variabili di input e l'etichetta, rendendo così possibile la previsione dell'etichetta anche per nuove istanze future, per le quali sono disponibili le sole variabili di input (previsori), mentre l'etichetta deve essere ancora osservata in quanto variabile di outcome. In questo senso, il termine *supervisionato* indica proprio l'apprendimento sulla base dell'informazione che è già stata osservata e disponibile. Nel nostro caso, l'etichetta binaria rappresenterà, evidentemente, il verificarsi oppure il non verificarsi dell'abbandono degli studi universitari. Con una codifica opportuna, possiamo inserire il verificarsi/non verificarsi di tale evento a livello individuale all'interno di un algoritmo di classificazione.

Le tecniche di apprendimento già utilizzate in letteratura la previsione dell'abbandono sono molteplici[12]. A titolo di esempio, possiamo citare: regressione logistica[13], algoritmi ad albero di tipo CART[14,15] (Classification and Regression Trees), algoritmo Naïve Bayes[16], Support Vector Machines[17] (SVM, macchine a supporto vettoriale), Artificial Neural Networks[18,19] (ANN, reti neurali artificiali), ed altre tecniche per le quali rimandiamo alla letteratura specifica[20]. Va tuttavia opportunamente sottolineato che non esiste un algoritmo di apprendimento che,

[8] Baker, Yacef, 2009, 3-17.
[9] Koedinger *et al.*, 2015, 333-353.
[10] Calvet Liñán, Juan Pérez, 2015, 98-112.
[11] Mitchell, 1997, chapter 6.
[12] Kumar *et al.*, 2017, 8-19.
[13] Willging, Johnson, 2004, 105-118.
[14] Dekker *et al.*, 2009, 41-50.
[15] Kumar, Pal, 2011, 63-69.
[16] Şara *et al.*, 2015, 319-324.
[17] Tekin, 2014, 207-226.
[18] Rios *et al.*, 2013, 289-294.
[19] Teshnizi, Ayatollahi, 2015, 296-300.
[20] Marquez-Vera *et al.*, 2013, 107-124.

uniformemente in ogni dominio applicativo, ottiene sistematicamente le performance previsive più elevate. Il confronto deve essere, invece, di natura essenzialmente empirica, sulla base dei metodi che saranno brevemente riassunti nella Sezione 3.4. Per questo motivo, gli algoritmi che confrontiamo in questo lavoro non sono stati selezionati in base ad un criterio decisionale ben definito, ma sono il risultato di una scelta 'ragionevole', che deve essere in ogni caso valutata a posteriori sulla base di un criterio puramente empirico.

Sulla base di queste considerazioni introduttive di carattere generale, nella Sezione 2 descriveremo le caratteristiche dei dati utilizzati in questo studio, riguardanti un sottoinsieme degli studenti iscritti presso l'Università di Bari Aldo Moro. La Sezione 3 contiene una breve introduzione agli algoritmi di classificazione utilizzati per la previsione degli abbandoni. La Sezione 4 riporta, in dettaglio, i risultati ottenuti. Infine, la Sezione 5 riassume brevemente l'oggetto e risultati di questo lavoro, individuando nuovi obiettivi di ricerca che potranno essere sviluppati in futuro.

2. I dati utilizzati in questo lavoro sono stati estratti dall'Osservatorio Studenti-Didattica del Miur Cineca. Tale Osservatorio è specificatamente riservato agli Atenei per la comunicazione degli eventi al Sistema Anagrafe Nazionale Studenti, che è a sua volta costituito da un vasto archivio amministrativo in cui vengono registrati gli iscritti al Sistema Universitario italiano. L'evento considerato (*variabile di outcome*) è l'uscita dal sistema per gli studenti iscritti presso l'Università di Bari Aldo Moro, negli anni dal 2013 al 2016. Le possibili condizioni all'uscita contemplate dal sistema sono: L ≡ Laurea, R ≡ Rinuncia, M ≡ Decesso: le istanze recanti il codice di uscita M, in numero molto esiguo, sono state preventivamente rimosse dal database. In realtà, il sistema contempla altre possibilità di uscita dal sistema (come ad esempio il codice 'D', associato alla decadenza dagli studi, oppure il codice 'I' associato alla rinuncia implicita agli studi). Tuttavia, queste eventualità non si sono mai verificate tra le istanze ricadenti nella finestra di osservazione presa in considerazione. Inoltre, i trasferimenti ad altra sede non sono presi in considerazione in questo lavoro, in quanto essi non sono considerati come uscite dal sistema universitario nazionale. Per ciascuna istanza, le variabili di profilazione messe a disposizione nel database dal sistema per ciascuna istanza, e considerate come variabili di input, saranno indicate in dettaglio nella Sezione 4. Nel seguito di questo paragrafo ci limiteremo ad una breve analisi delle variabili più importanti, che possa servire a cogliere le tendenze di fondo del fenomeno che stiamo studiando in questo lavoro, così come si sono manifestate per l'Università degli Studi di Bari.

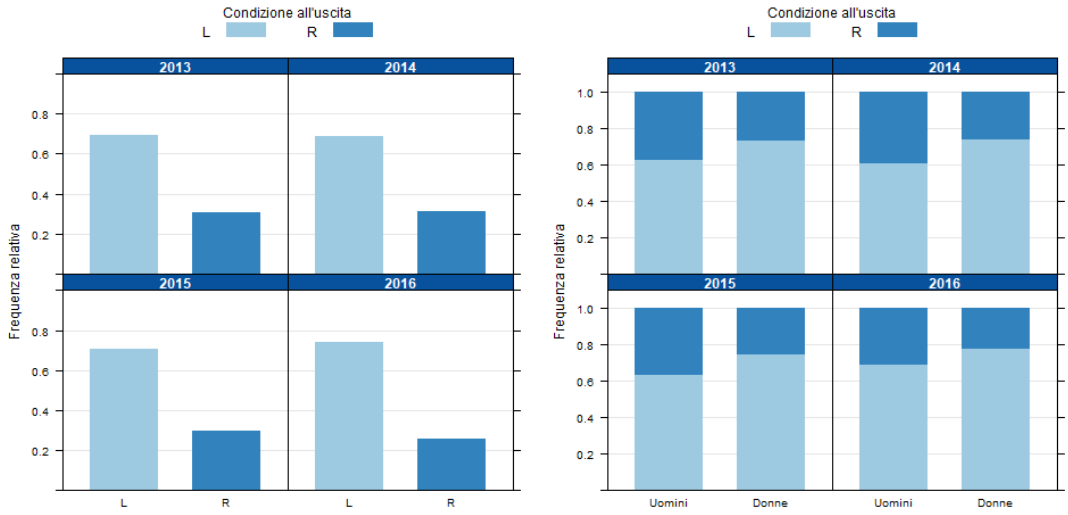

Figura 1. Sinistra: distribuzione delle frequenze relative per anno, suddivisa in base alla condizione all'uscita dal sistema universitario (L ≡ Laurea, R ≡ Rinuncia agli studi). Destra: distribuzione percentuale per anno di uscita e per sesso, suddivisa in base alla condizione all'uscita dal sistema universitario (L ≡ Laurea, R ≡ Rinuncia agli studi). Fonte: Osservatorio Studenti-Didattica del Miur-Cineca.

Analizziamo pertanto, a livello descrittivo, alcune caratteristiche della distribuzione di frequenza delle condizioni all'uscita, stratificata in base ad una o più variabili di interesse. La Figura 1 riporta la distribuzione delle frequenze relative per anno (pannello di sinistra) delle condizioni all'uscita: la frequenza percentuale dei rinunciatari passa dal 30.7% del totale nel 2013, al 25.7% nel 2016, segnando una lieve inversione di tendenza che, tuttavia, dovrà consolidarsi ed essere confermata negli anni seguenti. Il pannello di destra della Figura 1 riporta invece la distribuzione percentuale per anno ed ulteriormente stratificata in base al sesso. La percentuale dei rinunciatari appare decisamente più alta tra gli uomini, ed anche in questo caso il trend indicante una lieve riduzione degli abbandoni è confermato per entrambi i sessi (per gli uomini: dal 37.4% del 2013 al 30.9% del 2016; per le donne: dal 26.5% del 2013 al 22.6% del 2016).

Altrettanto interessante è l'analisi dell'età all'uscita dal sistema universitario. In generale, la distribuzione degli studenti laureati presenta una evidente asimmetria positiva (dati espressi in anni e frazioni di anno: min = 20.6, Q_1 = 23.5, mediana = 25.1, media = 26.3, Q_3 = 27.4, max = 77.1). L'età media alla laurea risulta essere di 26.3 anni, e il 25% degli studenti consegue la laurea ad una età superiore a 27.4 anni, fino ad un massimo di 77.1 anni. La distribuzione degli studenti che rinunciano agli studi ha sostanzialmente la stessa forma anche se, ovviamente, le misure di sintesi sono alquanto differenti nella parte sinistra della distribuzione (min = 18.1, Q_1 = 20.2, mediana = 21.4, media = 23.3, Q_3 = 24.4, max = 72.6). L'età media della rinuncia agli

studi è, pertanto, 23.3 anni, ma il 25% degli studenti rinuncia ad una età uguale od inferiore a 20.2 anni, confermando quindi l'importanza drammatica che il fenomeno degli abbandoni riveste nel passaggio dal primo al secondo anno dei CdS.

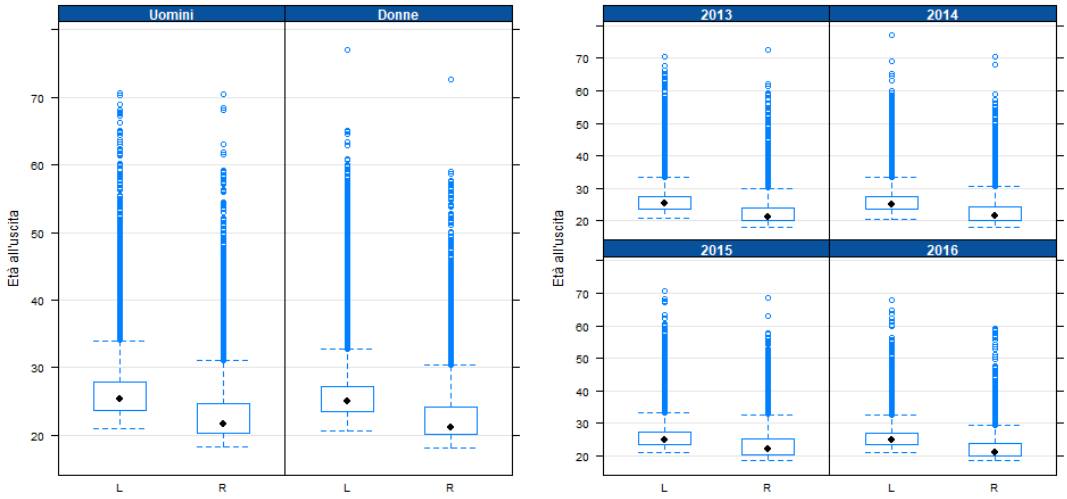

Figura 2. Sinistra: distribuzione dell'età all'uscita sistema universitario, suddivisa in base al sesso e alla condizione all'uscita dal sistema universitario (L ≡ Laurea, R ≡ Rinuncia agli studi). Destra: distribuzione dell'età all'uscita dal sistema universitario, suddivisa in base all'anno di uscita e alla condizione di uscita dal sistema dal sistema universitario (L ≡ Laurea, R ≡ Rinuncia agli studi). Fonte: Osservatorio Studenti-Didattica del Miur-Cineca.

Nella Figura 2 (pannello sinistro) riportiamo gli stessi dati disaggregati per sesso: non è evidente una sostanziale differenza, tra i due sessi, per quanto attiene l'età alla laurea (media = 26.7 anni per gli uomini, media = 26.1 anni per le donne). Invece, anche tenendo conto della suddivisione tra i due sessi, l'età media della rinuncia agli studi si abbassa drasticamente per entrambi i sessi (media = 23.5 anni per gli uomini, media = 23.1 anni per le donne, con il 25% degli uomini che rinunciano ad una età inferiore od uguale a 20.3 anni, e il 25% delle donne che rinunciano ad una età inferiore od uguale a 20.1 anni). Come evidente conseguenza di questi dati, possiamo affermare che l'età alla quale avviene l'abbandono degli studi non sembra essere legata in modo significativo al sesso dello studente. La stessa dinamica è assolutamente confermata anche dal punto di vista temporale se si guarda l'altro pannello della Figura 2, nel quale la distribuzione dell'età all'uscita dal sistema universitario è stratificata per il sesso, la condizione all'uscita e l'anno all'uscita.

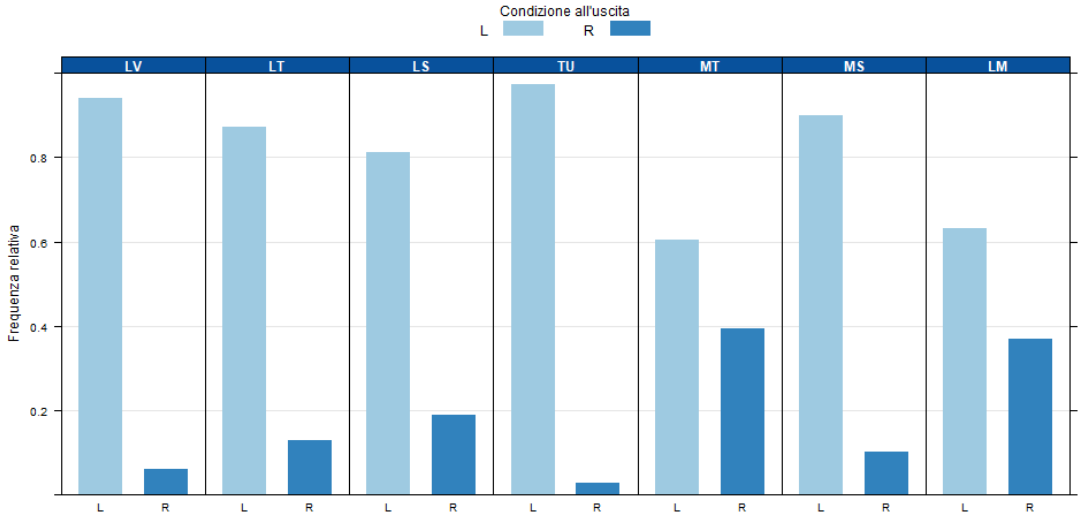

Figura 3: Distribuzione percentuale suddivisa in base alla condizione all'uscita dal sistema universitario nel periodo 2013-2016 (L ≡ Laurea, R ≡ Rinuncia agli studi) e alla tipologia del Corso di Studio (LV ≡ Laurea Vecchio Ordinamento, LT ≡ Laurea Triennale DM 509/99, LS ≡ Laurea Specialistica DM 509/99, TU ≡ Laurea a Ciclo Unico DM 509/99, MT ≡ Laurea Primo Livello DM 270/04, MS ≡ Laurea Magistrale DM 270/04, LM ≡ Laurea Magistrale a Ciclo Unico DM 270/04). Fonte: Osservatorio Studenti-Didattica del Miur-Cineca.

La Figura 3 riporta la distribuzione delle frequenze relative della condizione all'uscita (L ≡ Laurea, R ≡ Rinuncia), stratificata per la tipologia di Corso di Studio (CdS). È interessante osservare il dato molto interessante (in positivo) relativo agli abbandoni nel periodo considerato, osservato con i CdS appartenenti ai vecchi ordinamenti quadriennali (LV, lauree 93.9% contro il 6.1% di abbandoni). Tuttavia, questo dato va preso con evidente cautela: gli studenti ancora iscritti presso i vecchi ordinamenti quadriennali (che sono ad esaurimento ormai da molti anni), sono quasi certamente fortemente motivati a portare a termine il loro CdS. Il dato migliore lo si osserva invece con i CdS specialistici a ciclo unico del vecchio ordinamento ex DM 509/99 (TU, 97.2% di laureati contro il 2.9% di abbandoni nel periodo considerato). Anche in questo caso la motivazione gioca un ruolo fondamentale: la maggior parte degli studenti iscritti a CdS inquadrati in questo ordinamento sono studenti del Corso di Laurea a ciclo unico in Medicina. È quindi evidente che anche in questo caso gioca un ruolo la volontà di voler portare a termine gli studi in ogni caso: gli abbandoni osservati non sono della stessa natura degli abbandoni 'giovanili', che si verificano tra il primo e il secondo anno di corso, ma sono dovuti al verificarsi di circostanze che rendono improponibile la prosecuzione degli studi, nonostante la motivazione.

Per quanto riguarda le altre tipologie di CdS, gli abbandoni sono certamente più elevati in percentuale. Infatti, per i vecchi CdS Triennali ex DM 509/99 (LT) abbiamo

l'87.2% di laureati contro il 12.8% di abbandoni, mentre per i nuovi CdS di primo livello ex DM 270/04 (MT) rileviamo il 60.5% di laureati contro il 39.5% di abbandoni (nel periodo considerato). Dobbiamo tuttavia tener presente che i vecchi CdS Triennali ex DM 509/99 sono ad esaurimento, e quindi assume poca rilevanza il fenomeno degli abbandoni precoci, che è quello che ha maggior peso in valore assoluto e percentuale. Invece, decisamente alta appare la percentuale di abbandoni nei nuovi CdS di primo livello, per i quali il fenomeno degli abbandoni 'giovanili' (quelli che si verificano tra il primo e il secondo anno) gioca un ruolo certamente importante. Una dinamica leggermente differente la si osserva confrontando la Laurea Specialistica ex DM 509/99 (LS) con la Laurea Magistrale ex DM 270/04 (MS). Per la prima tipologia di CdS abbiamo l'81.2% di laureati contro il 18.8% di abbandoni, mentre per la seconda tipologia abbiamo il 90.0% di laureati contro il 10.0% di abbandoni. In entrambi i casi, l'età media elevata degli studenti e la loro motivazione a completare la parte conclusiva del loro percorso di studio si riflettono nella percentuale relativamente bassa di abbandoni.

Infine, è interessante osservare che per i CdS inquadrati tra le Lauree Magistrali a ciclo unico ex DM 270/04 (LM) si osserva, di nuovo, una percentuale di abbandoni molto elevata, e simile a quella della Lauree Triennali di primi livello (63.2% di laureati contro il 36.8% di abbandoni). Per questa tipologia di CdS, gli abbandoni tra il primo e il secondo anno giocano, ovviamente, un ruolo importante, reso ancora più rilevante dall'incertezza sul futuro derivante dall'iscrizione ad un CdS che, nella migliore delle ipotesi, impegnerà lo studente per almeno cinque anni.

Questi dati, presi nel loro complesso, sembrano quindi confermare l'importanza del fenomeno dell'abbandono precoce, che sebbene in lieve calo negli anni più recenti, ha ancora un peso quantitativamente molto elevato (sul totale degli studenti che sono all'interno del Sistema Universitario). Sebbene l'abbandono degli studi universitari, visto come fenomeno complessivo, si basi spesso su una decisione individuale, è evidente che ci sono delle regolarità che possono essere individuate ed utilizzate (sulla base delle tecnologie che abbiamo brevemente descritto nella Sezione introduttiva) per individuare in anticipo gli studenti che più sono a rischio, prima che la decisione di porre fine agli studi universitari sia effettivamente messa in atto.

3. Nei paragrafi che seguono, esporremo le tecniche di classificazione utilizzate per apprendere la probabilità di abbandono in funzione del vettore di caratteristiche (feature vector) associato a ciascuno studente, in modo da poter classificare istanze future prima che l'evento di interesse (abbandono degli studi universitari) si sia eventualmente verificato. Per fissare la notazione, C_k indicherà l'etichetta binaria associata a ciascuna istanza (ossia $k \in \{0,1\}$, con $k=1$ indicante la *positive class*, e cioè l'abbandono degli studi universitari); invece, $\mathbf{x} = (x_1,...,x_s) \in R^s$ denoterà il vettore di caratteristiche (variabili di input) associato a ciascuna istanza.

3.1. L'algoritmo *Naïve Bayes* (NB) è un tipico classificatore generativo[21], basato sull'apprendimento delle probabilità congiunte $p(\mathbf{x}, C_k) = p(\mathbf{x} \mid C_k)\pi(C_k)$ e sul teorema di Bayes per calcolare, a partire da queste, le probabilità a posteriori $p(C_k \mid \mathbf{x})$. Le stime di tali probabilità sono utilizzate direttamente per prevedere le etichette su istanze future, assegnando uno studente a quella classe in corrispondenza della quale la probabilità a posteriori è la più elevata possibile (classificatore MAP, *maximum a posteriori*).

Per esporne brevemente le caratteristiche, faremo l'ipotesi preliminare che le tutte le *s* variabili che costituiscono il vettore di input siano discrete, ossia che possano assumere un numero finito di valori. Più avanti, vedremo come questa restrizione possa essere agevolmente superata. Utilizzando il teorema di Bayes[22] abbiamo l'espressione delle probabilità a posteriori:

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\pi(C_k)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid C_k)\pi(C_k). \tag{1}$$

La quantità al denominatore della (1) prende il nome di *evidenza* o *verosimiglianza marginale*, ed è definita da:

$$p(\mathbf{x}) = \sum_{k=0,1} p(\mathbf{x} \mid C_k)\pi(C_k). \tag{2}$$

Essa rappresenta, evidentemente, la probabilità di verificarsi del vettore di variabili di input $\mathbf{x}$, senza avere informazione di alcun genere sulla relativa etichetta. Generalmente, la (2) è difficile da calcolare esplicitamente: tuttavia, poiché il nostro obiettivo è quello di massimizzare la (1) rispetto all'etichetta, possiamo trascurarne l'espressione, e concentrarci sulla massimizzazione del solo denominatore che appare nell'espressione a destra dell'uguaglianza garantita dal teorema di Bayes.

Poiché il vettore $\mathbf{x}$ ha generalmente dimensionalità elevata, la stima diretta delle probabilità multivariate $p(\mathbf{x} \mid C_K)$ tramite un dataset di training è spesso problematica, a causa della sparsità intrinseca dei dati. Infatti, molte delle possibili combinazioni delle variabili di input non appariranno affatto nel training set, così come la maggior parte delle combinazioni comparirà in un piccolo numero di istanze, rendendo le stime delle probabili multivariate fortemente instabili. Per questo, facciamo la seguente fondamentale ipotesi (*class conditional independence*):

[21] Ng, Jordan, 2002, 841-848.
[22] Hastie *et al.*, 2009, 21.

$$p(\mathbf{x} \mid C_k) \equiv \prod_{j=1}^{s} p(x_j \mid C_k), \tag{3}$$

ove la (3) non segue dalla struttura del problema, ma è parte integrante dell'algoritmo, e mira a sostituire la stima della probabilità condizionale multivariata che appare nel membro sinistro della (3), con quelle di s probabilità univariate (evidentemente più stabili, e meno condizionate dal problema della sparsità).

Sotto queste ipotesi, il classificatore teorico ha la seguente espressione:

$$\gamma(\mathbf{x}) = \arg\max_{k=0,1} p(\mathbf{x} \mid C_k)\pi\left(C_k\right), \tag{4}$$

del quale sono ben note le proprietà di ottimalità sotto una funzione di perdita 0/1[23]. Assegnato un dataset di training D nel quale compaiono N istanze ad informazione completa (per le quali abbiamo a disposizione tanto il vettore di input quanto l'etichetta), le stime di massima verosimiglianza delle probabilità a priori sono:

$$\hat{\pi}(C_k) = \frac{\#D\left[C_k\right]}{|D|}, \quad k = 0,1. \tag{5}$$

dove la scrittura $\# D\left[C_k\right]$ indica il numero di istanze nel training set la cui etichetta è C_k. Con la stessa convenzione, la stima di massima verosimiglianza di $p(x_j \mid C_k)$ ha espressione:

$$\hat{p}(x_j \mid C_k) = \frac{\#D\left[C_k \wedge x_j\right]}{\#D\left[C_k\right]}, \quad j = 1,\ldots,s. \tag{6}$$

Stante l'espressione delle stime di massima verosimiglianza, il classificatore empirico per una nuova istanza di test il cui vettore di input è $\mathbf{x}^{new}$ ha la seguente espressione:

$$\hat{\gamma}\left(\mathbf{x}^{new}\right) = \arg\max_{k=0,1} \prod_{j=1}^{s} \hat{p}(x_j^{new} \mid C_k)\hat{\pi}(C_k). \tag{7}$$

Naturalmente, bisogna tenere presente che il vettore di input $\mathbf{x}^{new}$ potrebbe non occorrere mai all'interno della classe k. In questo caso, dalla (6) consegue che la corrispondente istanza di test ha probabilità a posteriori nulla per l'etichetta C_k, e con probabilità 1 non verrà assegnata a tale classe. Se lo stesso vettore di input non si

[23] Hastie *et al.*, 2009, 22.

realizza mai, almeno una volta, in entrambe le classi, l'algoritmo NB non ha possibilità di funzionare per quella particolare istanza di test. In altre parole, non abbiamo alcuna possibilità di decidere sull'etichetta da attribuire. In questo caso, per attenuare la portata di questo problema, le stime di massima verosimiglianza (6) vengono lisciate nel modo seguente:

$$\hat{p}_\lambda(x_j \mid C_k) = \frac{\#D\left[C_k \wedge x_j\right] + \lambda}{\#D\left[C_k\right] + \lambda J}, \quad j = 1,...,s, \tag{8}$$

dove J è il numero di valori distinti che la variabile di input x_j può assumere, mentre λ è un parametro che governa l'ampiezza dello smoothing, e che deve essere settato in modo empirico dall'utente.

Se rilassiamo l'ipotesi che le variabili di input siano continue, la più comune variazione è l'algoritmo *Naïve Bayes gaussiano* (GNB, Gaussian Naïve Bayes), che funziona sotto l'ipotesi che:

$$f(x_j \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left\{-\frac{1}{2\sigma_{jk}^2}(x_j - \mu_{jk})^2\right\}, \tag{9}$$

ossia che la verosimiglianza condizionale in ciascuna classe sia una v.a. continua gaussiana, con parametri che nel caso più generale dipendono anche dall'etichetta di classe k, sebbene alcune restrizioni siano possibili, in base alle quali la varianza condizionale viene semplificata come $\sigma_{jk}^2 \equiv \sigma_j^2$ (ossia costante all'interno delle classi), ovvero $\sigma_{jk}^2 \equiv \sigma^2$ (ossia le variabili di input sono omoschedastiche, con variabilità che non varia neppure tra le classi). Nel caso più generale i parametri da stimare sono $2s \times 2$, e le relative stime di massima verosimiglianza hanno la seguente espressione:

$$\begin{aligned} \hat{\mu}_{jk} &= \frac{1}{\sum_{i=1}^{N} I\left[y^{(i)} = C_k\right]} \sum_{i=1}^{N} x_j^{(i)} I\left[y^{(i)} = C_k\right] \\ \hat{\sigma}_{jk}^2 &= \frac{1}{\sum_{i=1}^{N} I\left[y^{(i)} = C_k\right]} \sum_{i=1}^{N} (x_j^{(i)} - \hat{\mu}_{jk})^2 I\left[y^{(i)} = C_k\right] \end{aligned} \tag{10}$$

dove $x_j^{(i)}$ indica il valore assunto dalla *j*-esima variabile di input sulla *i*-esima istanza di training, mentre $y^{(i)}$ indica l'etichetta nella *i*-esima istanza di training, e quindi $I\left[y^{(i)} = C_k\right] = 1$ quando l'etichetta della *i*-esima istanza è effettivamente C_k.

Utilizzando la verosimiglianza gaussiana (9), il classificatore empirico assume la seguente espressione:

$$\hat{\gamma}_G\left(\mathbf{x}^{new}\right) = \arg\max_{k=0,1} \prod_{j=1}^{s} N(x_j^{new} \mid \hat{\mu}_{jk}, \hat{\sigma}_{jk}^2)\hat{\pi}(C_k). \tag{11}$$

Esistono, in letteratura, molti risultati sulle performance empiriche del classificatore Naïve Bayes, che spesso si è dimostrato superiore ad altri classificatori in molti contesti. Sono anche disponibili alcuni risultati teorici di carattere generale sulle possibilità di apprendimento (e sui relativi limiti) di questo algoritmo[24]: tuttavia, tali risultati dipendono spesso da ipotesi specifiche, e sono di difficile applicabilità nel lavoro pratico.

3.2. La *regressione logistica* si contrappone all'algoritmo Naïve Bayes in quanto le probabilità a posteriori $p(C_k \mid \mathbf{x})$ sono apprese direttamente, senza passare dalla stima della distribuzione di probabilità congiunta $p(C_k, \mathbf{x})$: in questo senso, la regressione logistica fornisce un approccio discriminativo alla classificazione di istanze future. In questa sezione sarà utile scrivere direttamente le etichette in formato binario 0/1 (ossia $C_1 \equiv 1$ e $C_0 \equiv 0$). Con questa convenzione, la forma parametrica delle equazioni di regressione logistica è la seguente:

$$\begin{aligned} p(y^{(i)} = 0 \mid \mathbf{x}^{(i)}) &= \frac{1}{1 + \exp(w_0 + \sum_{j=1}^{s} w_j x_j^{(i)})} \\ p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}) &= \frac{\exp(w_0 + \sum_{j=1}^{s} w_j x_j^{(i)})}{1 + \exp(w_0 + \sum_{j=1}^{s} w_j x_j^{(i)})} \end{aligned}, \tag{12}$$

La seconda delle (12) segue implicitamente dalla prima, poiché le due probabilità devono sommare ad 1. Globalmente, le equazioni (12) definiscono un modello lineare generalizzato (GLM, Generalized Linear Models) con funzione di link logistica.

Una particolarità interessante del modello di regressione logistica sta nel fatto che esso è equivalente ad una funzione di classificazione lineare. Infatti, se vogliamo attribuire un'istanza futura a quella classe la cui probabilità a posteriori è massima, possiamo affermare, in modo del tutto equivalente, che all'istanza il cui vettore di input è $\mathbf{x}^{new}$ viene assegnata l'etichetta $C_1 \equiv 1$ se e solo se:

[24] Ng, Jordan, 2002, 841-848.

$$\frac{p(y^{new}=1\,|\,\mathbf{x}^{new})}{p(y^{new}=0\,|\,\mathbf{x}^{new})}>1, \tag{13}$$

e sostituendo nella (13) le espressioni (12) otteniamo:

$$\exp(w_0+\sum_{j=1}^{s}w_j x_j^{new})>1. \tag{14}$$

Prendendo il logaritmo di entrambi i membri, è evidente che la funzione di classificazione teorica associata alla regressione logistica è la seguente:

$$\gamma(\mathbf{x}^{new})=C_1 \text{ se e solo se } w_0+\sum_{j=1}^{s}w_j x_j^{new}>0. \tag{15}$$

La stima dei parametri del modello di regressione logistica è generalmente basata sulla massimizzazione della log-verosimiglianza condizionale sull'insieme di training, ossia:

$$\hat{\mathbf{w}}\leftarrow\arg\max_{\mathbf{w}}\sum_{i=1}^{N}\log p(y^{(i)}\,|\,\mathbf{x}^{(i)},\mathbf{w}), \tag{16}$$

dove $\mathbf{w}=(w_0,w_1,...,w_s)$ è il vettore dei parametri rispetto al quale avviene la massimizzazione. Nel caso di un'etichetta binaria, è ben noto che la (16) ha un'espressione particolarmente semplice[25], anche se purtroppo non è possibile massimizzarla analiticamente rispetto al vettore dei parametri. Poiché anche il gradiente e la matrice Hessiana associati alla log-verosimiglianza condizionale hanno una forma molto semplice, la soluzione generalmente adottata è quella di utilizzare un metodo basato sulla discesa del gradiente del secondo ordine, generalmente noto in letteratura come Iterative Reweighted Least Squares (IRLS).

Per il modello di regressione logistica si pone spesso, come per tutti i modelli discriminativi, il problema dell'overfitting, soprattutto nel caso in cui il vettore di input abbia dimensione elevata e i dati del training set siano sparsi. In linea di principio, le stime dei parametri ottenute mediante il metodo della massima verosimiglianza possono essere trattate all'interno di una interessante teoria asintotica, che permette di calcolare i *p*-values associati ad ogni elemento del vettore delle stime, nonché la significativa della riduzione della quota di variabilità spiegata da un modello nidificato, nel quale alcuni previsori presenti tra le variabili di input siano stati rimossi (la tecnica prende il nome di *analisi della devianza*, e valuta la significatività della riduzione nella log-verosimiglianza rispetto al modello saturo[26]).

[25] Hastie *et al.*, 2009, 120.
[26] Firth, 1990, 55-82.

Tuttavia, la teoria asintotica che abbiamo appena richiamato non è standard, e non coincide con la teoria asintotica valida per il rapporto generalizzato di verosimiglianze quando $N \to \infty$. Sono invece necessarie alcune condizioni aggiuntive per evitare che le approssimazioni asintotiche siano di cattiva qualità. Per aggirare queste difficoltà tecniche, una possibile soluzione è quella di regolarizzare il modello di regressione logistica introducendo una penalità che scoraggi valori 'grandi' degli elementi di $\mathbf{w}$, forzando una parte degli elementi del vettore ad essere 'prossimi' a zero (*shrinkage* dei parametri). Una scelta possibile (e tipica) è quella della regolarizzazione basata su una funzione di perdita quadratica, detta anche perdita L_2, in base alla quale il nuovo obiettivo da massimizzare è[27,28]:

$$\hat{\mathbf{w}} \leftarrow \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log p(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w}\|^2, \tag{17}$$

dove γ prende il nome di *parametro di regolarizzazione*: la sua interpretazione discende dal fatto che si può facilmente dimostrare che il problema di massimizzazione (17) è equivalente alla stima MAP del vettore $\mathbf{w}$, supponendo che questo sia stato dotato di una distribuzione a priori gaussiana con valore atteso nullo e varianza $1/\gamma$. Dunque, valori elevati di γ favoriscono un ammontare di shrinkage progressivamente maggiore. Anche il metodo della discesa del gradiente può essere facilmente adattato per tenere conto della penalità.

3.3. L'algoritmo di classificazione supervisionata che discutiamo in questo sottoparagrafo è un algoritmo discriminativo che, a differenza dei due che lo hanno preceduto nell'esposizione, non ha carattere puramente probabilistico, ma è altresì basato su considerazioni essenzialmente geometriche. In questa sezione le etichette, per convenienza, verranno ulteriormente rinominate in modo tale che $y^{(i)} \in \{-1,+1\}$, dove per convenzione +1 rappresenta anche in questo caso la positive class. Indichiamo inoltre con $X \subseteq R^s$ lo spazio delle variabili di input, e con Y l'insieme binario nel quale vivono le etichette. Con questa notazione, un classificatore teorico è evidentemente una applicazione $\gamma: X \to Y$. Nella forma più generale possibile, tale classificatore teorico dipende da un vettore di parametri $\boldsymbol{\alpha}$, ed ha espressione:

$$\gamma(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\alpha}). \tag{18}$$

Di questa funzione ne abbiamo già visto un esempio nella sezione precedente, dedicata alla classificazione basata sulla regressione logistica. Se indichiamo con l la funzione di perdita 0/1, allora essa assume valore 0 solo quando la *true label* e la

[27] Cessie, van Houwelingen, 1992, 191-201.
[28] Goeman *et al.*, 2016, https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

predicted label, ottenuta tramite $f(\mathbf{x},\boldsymbol{\alpha}),$ coincidono. Possiamo allora definire, nel modo seguente, il *rischio atteso* associato alla funzione di classificazione teorica γ[29]:

$$R(\boldsymbol{\alpha}) = \int l(f(\mathbf{x},\boldsymbol{\alpha}), y) dP(\mathbf{x}, y) \tag{19}$$

Evidentemente, la (19) misura la probabilità di errata classificazione per una generica coppia input-etichetta $(\mathbf{x}, y)$ non ancora osservata, e dipende dalla distribuzione di probabilità incognita $P(\mathbf{x}, y)$. Per capire meglio la connessione di questa quantità con l'algoritmo che andiamo a descrivere, esplicitiamo la forma dell'insieme di training, contenente le N coppie input-etichetta $D = \{(\mathbf{x}^{(i)}, y^{(i)});\ i = 1,...,N\}$, osservate su N istanze. Supponiamo, inizialmente, che i dati del training set siano linearmente separabili: la separabilità lineare è il caso più comune (ma anche quello meno realistico), ed implica che gli N vettori di input possano essere perfettamente separati, rispetto alle etichette, da un iperpiano affine s dimensionale. Nell'ipotesi che i dati di training siano separabili, esistono certamente due istanze di training, $(\mathbf{x}^{+}, +1)$ e $(\mathbf{x}^{-}, -1)$, una avente etichetta positiva e l'altra avente etichetta negativa per le quali è possibile trovare un iperpiano affine di equazione $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ (noto come *superficie di decisione*, definito dall'intercetta b e dal vettore di giacitura $\mathbf{w} = (w_1,...,w_s)$), tale che le due istanze in questione abbiano (entrambe) distanza euclidea minima da tale superficie di decisione[30]. Si noti che, per economia di notazione, abbiamo utilizzato la notazione basata sul prodotto scalare tra due vettori, $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x} = \sum_{j=1}^{s} w_j x_j^{(i)}$, per compattare l'espressione algebrica dell'iperpiano affine. Possiamo allora definire due iperpiani affini, noti rispettivamente come *margine positivo* e *margine negativo*, paralleli alla superficie di decisione e tali che:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \geq +1 \quad & \text{per } y^{(i)} = +1 \\ \langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \leq -1 \quad & \text{per } y^{(i)} = -1 \end{aligned}, \tag{20}$$

Ovviamente, le due espressioni contenute in (20) possono essere sintetizzate come:

$$y^{(i)} \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right) - 1 \geq 0, \quad i = 1,...,N. \tag{21}$$

Se d_+ e d_- rappresentano, rispettivamente, la distanza euclidea della superficie di decisione dal margine positivo e da quello negativo, ci poniamo l'obiettivo di

[29] James *et al.*, 2013, chapter 4.
[30] Liu, 2011, 113.

determinare i parametri $\boldsymbol{\alpha} = (b, \mathbf{w})$ in modo tale che il *margine geometrico* $\rho = d_+ + d_-$ sia il più grande possibile. Con semplici considerazioni geometriche è possibile verificare che tale richiesta equivale al seguente problema di ottimizzazione quadratica vincolato:

$$\mathbf{w}_{opt} \leftarrow \arg\min_{\mathbf{w}} \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2, \tag{22}$$

sottoposto all'insieme dei vincoli lineari rappresentati dalla (21). Utilizzando gli strumenti della programmazione quadratica[31], è possibile dimostrare che la soluzione esiste nel caso linearmente separabile ed è sparsa, nel senso che gli elementi di $\boldsymbol{\alpha}_{opt} = (b_{opt}, \mathbf{w}_{opt})$ che massimizzano il margine geometrico dipendono esclusivamente dagli N_s vettori di input del training set che ricadono esattamente sui margini, e che sono tutti e soli i vettori che verificano la condizione:

$$y^{(i)}\left(\left\langle \mathbf{w}_{opt}, \mathbf{x}^{(i)} \right\rangle + b_{opt}\right) - 1 = 0, \quad i = 1, ..., N_s. \tag{23}$$

I vettori che soddisfano la (23) prendono il nome di *vettori di supporto*. Una volta ottenuti i parametri ottimali, il classificatore empirico ha la seguente espressione:

$$\hat{\gamma}(\mathbf{x}^{new}) = f(\mathbf{x}^{new}, \boldsymbol{\alpha}_{opt}) = sign\left(\left\langle \mathbf{w}_{opt}, \mathbf{x}^{new} \right\rangle + b_{opt}\right), \tag{24}$$

dove per definizione $sign(z) = +1$ se $z > 0$, mentre $sign(z) = -1$ se $z < 0$. Se definiamo il rischio empirico sul training set come:

$$R_{emp}(\boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} l(f(\mathbf{x}^{(i)}, \boldsymbol{\alpha}), y^{(i)}), \tag{25}$$

evidentemente la (25) misura la percentuale di istanze incorrettamente classificate nel training. Per giustificare il classificatore che abbiamo appena introdotto, si può partire allora dalla seguente diseguaglianza, che vale con probabilità (pari almeno a) $1 - \eta$[32]:

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h\left[\ln\left(\frac{2N}{h}\right) + 1\right] - \ln\left(\frac{\eta}{4}\right)}{N}}, \tag{26}$$

[31] Liu, 2011, 114.
[32] Vapnik, 1999, 998-999.

dove h indica la dimensione di Vapnik-Chervonenkis (VC)[33] della classe di funzioni $f(\mathbf{x},\boldsymbol{\alpha})$. Quando tale funzione di classificazione teorica ha la stessa forma di quella utilizzata nella (24) per il classificatore empirico, si può dimostrare che:

$$h \leq \min\left(\left\lceil \frac{R^2}{\rho^2} \right\rceil, s\right)+1, \tag{27}$$

dove R indica il raggio della più piccola ipersfera contenente tutti i vettori di input. Ma allora, massimizzare il margine ρ equivale a minimizzare la dimensione di VC. In questo caso $\boldsymbol{\alpha}=\boldsymbol{\alpha}_{opt}$, e il corrispondente errore di training è nullo nel caso di un dataset linearmente separabile. Dunque, attraverso la superficie di decisione di massimo margine stiamo minimizzando implicitamente la probabilità di incorretta classificazione per una istanza non osservata, senza che sia necessaria alcuna informazione sulla distribuzione di probabilità congiunta $P(\mathbf{x}, y)$.

Naturalmente, quando il dataset di training non è linearmente separabile, tutta la teoria che abbiamo esposto sin qui non è più valida. Una soluzione possibile è quella di rilassare i vincoli (20) nel modo seguente[34]:

$$\begin{aligned} \left\langle \mathbf{w}, \mathbf{x}^{(i)} \right\rangle + b \geq +1 - \xi^{(i)} \quad & \text{per } y^{(i)} = +1 \\ \left\langle \mathbf{w}, \mathbf{x}^{(i)} \right\rangle + b \leq -1 + \xi^{(i)} \quad & \text{per } y^{(i)} = -1 \end{aligned}, \tag{28}$$

dove le quantità $\xi_i \geq 0$ prendono il nome di *variabili slack*. Esse sono evidentemente comprese tra 0 ed 1 solo quando una data superficie di decisione classifica correttamente la corrispondente istanza di training (eventualmente questa potrebbe ricadere, a differenza del caso precedente, all'interno del margine), mentre sono maggiori di 1 per quelle istanze di training che sono classificate incorrettamente. Dunque, $\sum_{i=1}^{N} \xi_i$ è un maggiorante del numero di istanze di training incorrettamente classificate, ed è naturale penalizzare per gli errori di classificazione considerando la seguente nuova funzione obiettivo:

$$\mathbf{w}_{opt} \leftarrow \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{N} \xi_i\right), \tag{29}$$

[33] Bartlett, 2001, 1-16.
[34] Manning *et al.*, 2008, 300.

con C parametro di costo che governa l'ammontare della penalizzazione. Anche in questo caso, abbiamo una formulazione opportuna di questo problema di ottimizzazione, che permette di determinare i parametri della superficie di decisione lineare ottimale.

3.4. Un modo comune di valutare empiricamente l'accuratezza dei classificatori che abbiamo descritto nella sezione precedenti è quello di suddividere, in modo causale, l'insieme dei dati a disposizione in due sottoinsiemi che non hanno istanze in comune, ossia *l'insieme di training* e quello di *test*. Il dataset di training viene utilizzato esclusivamente per l'apprendimento del classificatore, mentre il dataset di test è utilizzato per valutare l'accuratezza empirica, surrogando la disponibilità di istanze future per le quali l'etichetta non è nota. Infatti, la valutazione del rischio empirico sul training set porterebbe invariabilmente ad una sottovalutazione del rischio atteso (19), così come è dimostrato dalla (26) nel caso delle SVM. Anzi, l'ottimizzazione del rischio empirico sul training set porta facilmente a fenomeni di overfitting, in quanto spinge il classificatore ad adattarsi il più possibile ai dati di apprendimento, con la conseguenza di una performance previsiva scadente su dati futuri che dovessero mostrare deviazioni anche minime dal comportamento valido per i dati di apprendimento[35].

Invece, valutare l'accuratezza empirica su dati futuri indipendenti dal dataset di training protegge dalla distorsione negativa nella stima del rischio atteso. Una volta che le etichette previste nel test set dall'algoritmo di classificazione e le relative true label siano state disposte in una matrice di confusione 2×2, possiamo ovviamente calcolare il rischio empirico nella seguente forma:

$$\text{accuratezza} = \frac{TP + TN + FP + FN}{N_{\text{test}}}. \tag{30}$$

Nella forma indicata dalla (30), il rischio empirico calcolato sul test prende appunto il nome di *accuratezza* (dove, come è ben noto TP sta per 'true positives', e rappresenta il numero di istanze della positive class che sono state correttamente classificate, e così via). Esistono tuttavia altre metriche che sono più espressive per il tipo di problema che stiamo studiando, e in particolare[36]:

$$\begin{aligned} \text{sensibilità} &= \frac{TP}{TP + FN} \\ \text{specificità} &= \frac{TN}{TN + FP} \end{aligned}. \tag{31}$$

[35] Hastie *et al.*, 2009, 219.
[36] Parikh *et al.*, 2008, 45-40.

La *sensibilità*, nota anche come *recall* o *true positive rate* (TPR) misura la proporzione dell'istanze della positive class che sono correttamente classificate, ossia la proporzione degli studenti per i quali viene previsto l'abbandono degli studi, e che effettivamente abbandoneranno il corso al quale si sono iscritti. Invece, la *specificità* è nota anche come *true negative rate* (TNR), e misura la percentuale di istanze della negative class che sono correttamente classificate. Dunque, la specificità misura la capacità del classificatore di individuare correttamente gli studenti che proseguiranno il corso di studi fino alla laurea. Ottimizzare sul test set per la sensibilità o la specificità significa evidentemente perseguire obiettivi differenti, ed esiste un trade-off tra le due misure, nel senso che ottimizzare per una delle due significa, generalmente, ridurre il valore dell'altra[37].

Qualunque sia la metrica utilizzata per valutare la performance sul test, esiste un altro problema del quale dobbiamo preoccuparci. Normalmente, l'ottimizzazione sul test set viene effettuata non solo per poter confrontare classificatori differenti, bensì anche per poter scegliere i parametri liberi di un certo algoritmo. Per esempio, abbiamo visto che una SVM per dati non linearmente separabili dipende in modo critico dalla scelta del parametro di costo C. Anche in questo caso, una volta scelta la metrica da utilizzare sul test, il parametro libero viene settato a quel valore per il quale la metrica scelta assume il valore più elevato sul test set. Esiste tuttavia il rischio che questo processo di ottimizzazione finisca per dipendere troppo dalle caratteristiche del test set, e che quindi il rischio empirico sia, anche in questo caso, una sottostima del rischio atteso. Una soluzione generalmente accettata è quella di dividere l'insieme dei dati a disposizione in tre sottoinsiemi, training-test-validazione: i sottoinsiemi di training e di test sono utilizzati nel modo che abbiamo fin qui descritto, mentre il sottoinsieme di validazione viene utilizzato per rivalutare l'accuratezza rispetto al classificatore ottimale scelto mediante il test set. In questo possiamo ulteriormente surrogare la disponibilità di nuovi dati futuri che non sono entrati in alcun modo nel processo di training/test, e possiamo valutare se l'accuratezza che abbiamo trovato alla fine della fase di test è sostanzialmente confermata, oppure si osserva un drastico decremento.

Non esiste una regola univoca per la suddivisione dell'insieme dei dati nei tre sottoinsiemi training/test/validazione: una scelta puramente empirica spesso utilizzata in letteratura[38], e che seguiremo in questo lavoro, è 70%-20%-10%.

4. L'intero dataset utilizzato nello studio copre, come abbiamo detto, il periodo tra il 2013 e il 2016, e contiene in tutto $N = 41964$ istanze (studenti). Non abbiamo effettuato né una analisi anno per anno, né una analisi per coorte (in quanto il dato relativo alla coorte di immatricolazione non era disponibile al momento della scrittura

[37] Márquez-Vera *et al.*, 2016, 107-124.
[38] Hastie et al., 2009, Chapter 7.

del lavoro). Infatti, lo scopo implicito è quello di esplorare la performance generica di un insieme di algoritmi di classificazione supervisionata, rimandando a lavori futuri analisi più sofisticate e precise. Stante questa premessa, utilizzando le percentuali di splitting indicate alla fine della sezione precedente, le numerosità campionarie relative al processo di apprendimento/testing/validazione sono risultate essere: $N_{train} = 29374$, $N_{test} = 8393$, $N_{\text{validate}} = 4197$.

La scelta delle variabili utilizzate per la previsione non è stata fatta utilizzando metodi automatici, bensì scegliendo un set minimale costituito dalle seguenti sei variabili indipendenti:

1. ID NAZIONE NASCITA: variabile categoriale con due livelli: "Italia" ed "Estero";
2. REGIONE NASCITA: variabile categoriale con ovvio significato;
3. ID TIPO CDS: variabile categoriale che descrive il corso di studio intrapreso. I livelli di questa variabile categoriale sono quelli descritti nella didascalia della Figura 3;
4. REGIONE RESIDENZA: variabile categoriale con due livelli: "Italia" ed "Estero;
5. SESSO: variabile categoriale indicante il sesso dello studente;
6. ETÀ ALL'USCITA: età al momento del conseguimento del titolo di studio, ovvero dell'abbandono degli studi universitari. Sebbene questa variabile dipenda dall'evento che si vuol prevedere, essa non è un previsore perfetto del conseguimento del titolo di studio oppure dell'abbandono, e quindi può essere utilizzata per introdurre nel modello la probabilità di abbandono in funzione dell'età.

Tutte le variabili categoriali sono state codificate mediante variabili dummy basate su un livello scelto come riferimento[39]. Altre variabili categoriali, sebbene disponibili nel dataset originale, sono scartate poiché costituite da un numero di livelli troppo elevato (come, ad esempio, il comune di nascita o di residenza), che avrebbe portato ad una sovra-parametrizzazione eccessiva dei modelli di classificazione da apprendere. Inoltre, l'uso di un set così limitato di variabili indipendenti ha anche la funzione di 'stressare' l'uso dei modelli previsivi in una situazione di carenza di informazione disponibile.

Ciascun algoritmo di classificazione è stato appreso mediante il training set, facendo variare i parametri di controllo (solo per la regressione logistica semplice non abbiamo parametri di controllo a disposizione). Per ciascun valore prescelto di questi ultimi, abbiamo calcolato le metriche di accuratezza sul test set, ottimizzando per l'accuratezza e la sensibilità. Infine, dato il valore ottimale del parametro di controllo sul test set, abbiamo ulteriormente testato l'algoritmo, ricalcolando l'intero set di metriche sul validation set. Ciò allo scopo, come abbiamo già detto, di eliminare

[39] Agresti, 1990, 84.

l'eventuale influenza del particolare test set utilizzato nella scelta dei parametri, e di ritestare ulteriormente il modello appreso su un nuovo set di dati futuri indipendenti.

Regressione Logistica	**Acc.**	**TP**	**FP**	**TN**	**FN**	**Sens.**	**Spec.**
	0.77	757	224	5721	1691	0.31	0.96
Regress. Logistica L_2	**Acc.**	**TP**	**FP**	**TN**	**FN**	**Sens.**	**Spec.**
$\gamma = 0.01$	0.77	757	224	5721	1691	0.31	0.96
$\gamma = 0.1$	0.77	758	221	5724	1690	0.31	0.96
$\gamma = 1$	0.77	757	218	5727	1691	0.31	0.96
$\gamma = 10$	0.77	748	209	5736	1700	0.31	0.96
$\gamma = 100$	0.78	715	146	5799	1733	0.29	0.98
$\gamma = 1000$	0.78	644	2	5943	1804	0.26	1.00
$\gamma = 10000$	0.74	224	0	5945	2224	0.09	1.00
Regress. Logist. L_2							
$\gamma = 100$ (migliore acc.)	0.78	339	73	2918	867	0.28	0.98
$\gamma = 0.01$ (migliore sens.)	0.77	356	121	2870	850	0.30	0.96
Gaussian NB (Test)	**Acc.**	**TP**	**FP**	**TN**	**FN**	**Sens.**	**Spec.**
$\lambda = 0$	0.77	795	1653	5637	308	0.72	0.77
$\lambda = 1$	0.77	796	1652	5639	306	0.72	0.77
$\lambda = 2$	0.77	795	1653	5639	306	0.72	0.77
$\lambda = 3$	0.77	787	1661	5644	301	0.72	0.77
Gaussian NB							
$\lambda = 0$ (migl. acc. e sens.)	0.76	377	829	2821	170	0.69	0.77
SVM (Test)	**Acc.**	**TP**	**FP**	**TN**	**FN**	**Sens.**	**Spec.**
$C = 0.01$	0.80	818	1630	5932	13	0.98	0.78
$C = 0.1$	0.83	1111	1337	5863	82	0.93	0.81
$C = 1$	0.83	1148	1300	5848	97	0.92	0.82
$C = 10$	0.83	1153	1295	5848	97	0.92	0.82
$C = 100$	0.71	0	2448	5943	2	0.00	0.71
$C = 1000$	0.71	0	2448	5943	2	0.00	0.71
$C = 10000$	0.71	0	2448	5944	1	0.00	0.71
SVM (Validation)							
$C = 0.1$ (migliore acc.)	0.83	544	662	2956	35	0.94	0.82
$C = 0.01$ (migliore sens.)	0.80	391	815	2979	12	0.97	0.79

Tabella 1. Metriche di accuratezza previsiva relative ai classificatori utilizzati (SVM ≡ Support Vector Machines, NB ≡ Naïve Bayes). Per ciascun classificatore, le metriche sono state calcolate sul test set, facendo variare i parametri di controllo ove presenti. Solo in questo caso, le metriche sono state ricalcolate sul validation set, utilizzando i parametri di controllo ottimali che garantiscono, rispettivamente, la migliore accuratezza e la migliore sensibilità sul test set.

L'insieme completo dei risultati è riportato nella Tabella 1. La regressione logistica apre la tabella, poiché può essere evidentemente considerata come l'algoritmo di previsione 'baseline' con il quale gli altri devono confrontarsi. A dispetto di una accuratezza del 77%, la sensibilità risulta essere del 31% (ossia, tra gli studenti che abbandonano, solo tre su dieci vengono correttamente individuati). La bassa sensibilità è riflessa nel numero elevato di Falsi Negativi (FN), cioè tutti quegli studenti che abbandonano gli studi universitari senza essere individuati dall'algoritmo.

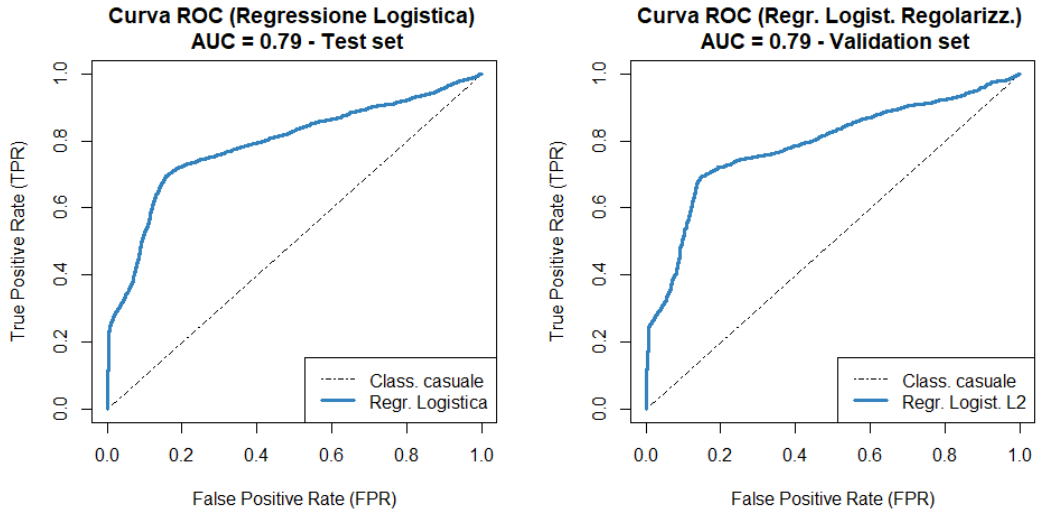

Figura 4. Sinistra: Curva ROC (≡ Receiver Operating Characteristic) per l'algoritmo di regressione logistica, calcolata sul test set. Destra: Curva ROC per l'algoritmo di regressione logistica regolarizzato, calcolata sul validation set, utilizzando il parametro di controllo $\gamma = 0.01$ che ottimizza la sensibilità sul test (si veda la Tabella 1). La linea tratteggiata corrisponde alla curva ROC del classificatore casuale.

Anche la regressione logistica regolarizzata non apporta un reale vantaggio: anzi, per valori elevati del parametro di regolarizzazione le prestazioni sul test set decadono nettamente. Questo fatto non deve sorprendere, dato la dimensione limitata del vettore delle variabili indipendenti, fatto che impedisce alla regressione logistica regolarizzata di esprimere un miglioramento nelle previsioni attraverso la riduzione di eventuali fenomeni di overfitting. Tuttavia, anche con questi semplici algoritmi, il vantaggio rispetto ad un classificatore casuale (che consiste nel classificare uno studente futuro come 'L' ovvero 'R' sulla base del lancio di una moneta non truccata) è testimoniato dalle relative curve ROC[40,41] (Receiver Operating Characteristics) riportate nella Figura 4, che in entrambi i casi si situano ben al di sopra della curva

[40] Liu, 2011, 83.
[41] Fawcett, 2006, 861-874.

ROC del classificatore casuale (linea tratteggiata), e corrispondono ad un AUC (Area Under the Curve) pari a 0.79 (l'AUC del classificatore casuale è 0.50, mentre quella del classificatore perfetto è 1.00).

Con l'algoritmo Gaussian NB otteniamo un netto miglioramento. Infatti, abbiamo una accuratezza del 76% e una sensibilità del 69% sul validation set (sette studenti rinunciatari su dieci sono correttamente identificati). Si noti come il numero di Falsi Negativi (FN) decade nettamente. Tuttavia, a ben guardare, il numero di Falsi Positivi (FP) cresce nettamente rispetto alla regressione logistica. Questo fatto è l'aspetto principale dell'inevitabile trade-off che esiste tra sensibilità e specificità. Tuttavia, dal nostro punto di vista, l'errore più grave è quello che si verifica quando non siamo in grado, in molti casi, di individuare uno studente che abbandonerà gli studi (errore di I° tipo), con conseguente riduzione della sensibilità. Invece, il declino nella specificità è dovuto al classificare erroneamente, in molti casi, come rinunciatari studenti che in realtà avrebbero portato a termine il Corso di Studi (errore di II° tipo). Ma, a ben vedere, quest'ultimo tipo di errore è nettamente meno grave poiché, nella peggiore delle ipotesi, porterà ad un maggiore utilizzo di risorse per seguire un maggior numero di studenti (risorse che, quindi, non saranno sprecate, bensì utilizzate per garantire servizi ad un numero di studenti più elevato).

La stessa tendenza la si osserva con l'algoritmo SVM. In questo caso, le prestazioni sul validation set sono davvero interessanti. Ottimizzando per la sensibilità, otteniamo una accuratezza dell'80% e una sensibilità del 97% (ossia quasi tutti gli studenti che abbandoneranno sono correttamente identificati), con un valore di specificità che si mantiene molto alto e pari al 79%. Anche in questo caso il numero di Falsi Positivi (FP) è elevato, sebbene leggermente minore del valore riscontrato con l'algoritmo Gaussian NB. Si noti che se ottimizziamo per l'accuratezza, il numero dei Falsi Positivi (FP) si riduce notevolmente, a scapito di una leggera riduzione della sensibilità, che scende al 94%.

5. I risultati riportati in questo lavoro sono indubbiamente incoraggianti, e mettono in evidenza l'utilità delle tecniche di apprendimento supervisionato nella previsione degli studenti che sono a rischio di abbandono degli studi universitari. Ovviamente, una serie di miglioramenti sono possibili: la prima ovvia considerazione è che la qualità dell'informazione disponibile è, spesso, molto più importante del particolare classificatore utilizzato. Sarà pertanto necessario sperimentare con un set di previsori più esteso di quello usato attualmente. Altre variabili esplicative delle quali potrebbe essere utile valutare la capacità previsiva sono: l'anno di immatricolazione (coorte di studio), il numero di crediti ottenuti fino alla conclusione del primo o del secondo anno di corso, la presenza o meno di un sostegno finanziario da parte delle istituzioni durante lo svolgimento degli studi. Altri attributi interessanti, che hanno dimostrato di possedere capacità previsiva, sono il voto di diploma

ottenuto al termine della scuola secondaria, l'interesse suscitato dai corsi seguiti e l'obbligo di frequenza[42].

Riguardo alle valutazioni di accuratezza previsiva che abbiamo presentato, una precisione è necessaria. Tutte le analisi sono state condotte sulla base di un singolo splitting training/test/validation. Ciò potrebbe ovviamente risultare limitativo rispetto a situazioni reali, e potrebbe essere causa di una sottovalutazione dell'errore di generalizzazione, con una conseguente riduzione della performance nel momento in cui dovessimo applicare il sistema previsivo a dati futuri. Esistono varie soluzioni proposte in letteratura per attenuare ulteriormente questo problema: quella concettualmente più semplice consiste nel ricorrere ad uno schema di tipo *k*-fold cross-validation, nel quale la stima dell'accuratezza è una media delle accuratezze ottenute su ciascun fold[43]. Si dimostra che, utilizzando una scelta appropriata del numero di fold, l'errore di generalizzazione può essere stimato in modo non distorto, e con maggior precisione rispetto alle stime calcolate su un singolo fold[44].

Pur con le limitazioni che abbiamo brevemente descritto, riteniamo che l'utilizzo delle tecniche di Machine Learning a problemi come quello che abbiamo trattato in questo lavoro possa ritenersi estremamente promettente ed utile. Una corretta individuazione precoce degli studenti a rischio di abbandono degli studi universitari è il primo passo per migliorare l'efficacia del nostro sistema di istruzione superiore, e promuovere un incremento nel numero di studenti che conseguono una laurea.

Riferimenti bibliografici

Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2016). Rapporto sullo Stato del Sistema Universitario e della Ricerca 2016 (Rapporto Integrale).

Agresti, A. (1990). *Categorical Data Analysis*, 2nd Edition. John Wiley & Sons. ISBN: 978-0-471-85301-1.

Baker, R.S.J.D. (2017). Data Mining for Education. To appear in: McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education* (3rd edition). Oxford, UK: Elsevier, in press.

Baker R.S.J.D., Yacef K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Bartlett, P. (2001). Statistical learning and VC theory. In *Tutorial Guide. ISCAS 2001. IEEE International Symposium on Circuits and Systems* (p. 4.2.1-4.2.16). IEEE. DOI: 10.1109/TUTCAS.2001.946954.

Belloc F., Maruotti A., Petrella, L. (2009). University drop-out: an Italian experience. *Higher Education*, *60(2)*, 127–138. DOI: 10.1007/s10734-009-9290-1.

[42] Márquez-Vera *et al.*, 2016, 107-124.
[43] Witten *et al.*, 2016, chapter 5.
[44] Hastie *et al.*, 2009, 241.

Bennett R. (2003). Determinants of Undergraduate Student Drop Out Rates in a University Business Studies Department. *Journal of Further and Higher Education*, *27(2)*, 123-141. DOI: 10.1080/030987703200065154.

Calvet Liñán L., Juan Pérez Á.A. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, *12*(3), 98-112. DOI: 10.7238/rusc.v12i3.2515.

Cessie S., van Houwelingen J.C., (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41(1)*, 191–201.

Comitato Nazionale per la Valutazione del Sistema Universitario. Undicesimo Rapporto sullo Stato del Sistema Universitario (2011). *Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR)*.

Dekker G., Pechenizkiy M., Vleeshouwers, J. (2009). Predicting Students Drop Out: A Case Study. In T. Barnes, M. C. Desmarais, C. Romero, & S. Ventura (Eds.), *EDM* (pp. 41–50). www.educationaldatamining.org.

Di Pietro G. (2004). The determinants of university dropout in Italy: A bivariate probit model with sample selection. *Applied Economics Letters*, *11*, 187-191.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. DOI: 10.1016/j.patrec.2005.10.010.

Firth D. (1990). Generalized Linear Models. *In Statistical Theory and Modelling: In Honour of Sir David Cox, FRS*, Eds. D V Hinkley, N Reid, E J Snell, pp. 55–82. London: Chapman and Hall.

Goeman J., Meijer R., Chaturvedi N. (2016). L_1 and L_2 Penalized Regression Models. Technical Document for CRAN.

Hastie T., Tibshirani R., Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York, NY: Springer New York. DOI: 10.1007/978-0-387-84858-7.

James G., Witten D., Hastie T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer New York. DOI: 10.1007/978-1-4614-7138-7.

Koedinger K.R., D'Mello S., McLaughlin E.A., Pardos Z.A., Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 333–353. DOI: 10.1002/wcs.1350.

Kumar B., Pal S. (2011). Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications*, *2*(6), 63-69. DOI: 10.14569/IJACSA.2011.020609.

Kumar M., Singh A. J., Handa D. (2017). Literature Survey on Educational Dropout Prediction. *International Journal of Education and Management Engineering*, *7*(2), 8–19. https://doi.org/10.5815/ijeme.2017.02.02.

Liu B. (2011). *Web data mining exploring hyperlinks, contents, and usage data, 2nd Edition*. Springer, Berlin.

Manning C.D., Raghavan P., Schütze H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Márquez-Vera C., Morales C.R., Soto S.V. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *8*(1), 7–14. DOI: 10.1109/RITA.2013.2244695.

Márquez-Vera C., Cano A., Romero C., Noaman A.Y.M., Mousa Fardoun H., Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

Mitchell T.M. (1997). *Machine Learning* (1st ed.). New York, NY, USA. McGraw-Hill, Inc.

Mohamad S. K., Tasir Z. (2013). Educational Data Mining: A Review. *Procedia - Social and Behavioral Sciences*, 97, 320–324. DOI: 10.1016/j.sbspro.2013.10.240.

Ng A.Y., Jordan, M.I. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 841–848). MIT Press.

Parikh R., Mathai A., Parikh S., Chandra Sekhar G., Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50. PMID: 18580002.

Rios G., Reyes N., Juárez M., Espitia E., Mosqueda J., Soria M. (2013). Predicting Early Students with High Risk to Drop Out of University using a Neural Network-Based Approach. in *ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology* (pp. 289-294). ISBN: 978-1-61208-283-7.

Şara N.B., Halland R., Igel C., Alstrup, S. (2015). High-school dropout prediction using machine learning: a Danish large-scale study. In M. Verleysen (Ed.), *Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 319-324).

Tekin A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research*, *54*, 207-226.

Teshnizi S., Ayatollahi S. (2015). A Comparison of Logistic Regression Model and Artificial Neural Networks in Predicting of Student's Academic Failure. *Acta Informatica Medica*, *23*(5), 296-300. DOI: 10.5455/aim.2015.23.296-300.

Vapnik V.N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. DOI: 10.1109/72.788640

Willging P.A., Johnson S.D. (2004). Factors that Influence Students' Decision to Dropout of Online Courses. *Journal of Asynchronous Learning Networks*, *8*(4), 105–118.

Witten I.H, Frank E., Hall M.A., Pal C. (2016). *Data Mining*, Fourth Edition. Morgan Kaufman. ISBN: 978-0-12804-291-5.