Capitolo 1

Introduzione

Dropout è un termine con il quale vengono indicati i ragazzi che decidono di abbandonare un percorso di studio prima della fine dello stesso. Il fenomeno del Dropout può avere ripercussioni molto importanti sulla vita dello studente ed è anche un fenomeno economicamente rilevante nella vita delle università. L'Alma Mater Studiorum - Università di Bologna ha raccolto i dati degli iscritti dal 2016 al 2018 per effettuare una ricerca relativa al fenomeno di abbandono degli studi. Lo studio ha lo scopo di andare a valutare se è possibile predire l'abbandono degli studi da parte di uno studente in modo tale da proporre agli studenti delle misure a sostegno di essi (come corsi di potenziamento) per aiutare i ragazzi ed evitare il fenomeno del Dropout.
Il lavoro proposto vuole valutare la possibilità di predire anticipatamente in vari momenti la possibilità che uno studente abbandoni gli studi. I momenti principali sui quali ci andremo a concentrare saranno tre:

1. Al momento dell'immatricolazione, non tenendo quindi conto degli esami svolti durante l'anno, ma considerando solo i dati personali ed economici del ragazzo ed eventuali OFA

2. Dopo la sessione invernale del primo anno, quindi considerando gli esami svolti al primo Marzo

3. Dopo il primo anno di corso, considerando quindi tutti gli esami svolti durante il primo anno di corso.

Questo ci permetterà di andare a valutare con che grado di accuratezza sarà possibile andare a predire la possibilità dell'abbandono da parte di uno studente durante il suo primo anno di studi, per cercare tramite delle misure di sostegno al ragazzo di scongiurare il fenomeno.
Abbiamo per i tre momenti precedentemente descritti utilizzato tre modelli di Machine Learning diversi: la *Regressione Logistica (LR)*, il *Random Forest (RF)* e la *Support*

Vector Machine (SVM). Abbiamo inoltre provato a combinare i risultati di questi tre modelli tramite lo *Stacking Algorithm.*

1.1 Lavori Correlati

Questo lavoro non è il primo che si pone come obiettivo quello di andare a valutare l'abbandono scolastico e non è il primo che cerca di farlo utilizzando il dataset offerto dall'Università di Bologna. Tra i più importanti lavori in questo senso troviamo sicuramente il lavoro di F. Del Bonifro et Al. [**?**], su cui il nostro lavoro si è fortemente basato e che utilizza RandomForest, SVM e LDA per andare a calcolare la possibilità che uno studente abbandoni o meno gli studi. Tramite diverse ricerche abbiamo trovato uno studio simile svolto all'Università Roma 3 [**?**]. In questo lavoro troviamo 3 insiemi di suddivisione delle features, il primo contiene tutte le carattteristiche sia accademiche che amministrative, il secondo contiene solo le features amministrative mentre il terzo aggrega le features amministrative insieme alle informazioni statistiche sulle carriere degli studenti, su cui vengono valutati due algoritmi basati sulle reti neurali(NB e kNN). Un lavoro che utilizza un approccio diverso dagli altri e si propone di andare a valutare il Dropout degli studenti a partire da informazioni delle emozioni provate dagli studenti e da altri fattori di emarginalizzazione economica, sociale e accademica è il lavoro proposto da Kadar et Al. [**?**].

1.2 Struttura del dataset

Il dataset contiene numerose informazioni che riassumiamo in questa tabella:

Campo	**Descrizione**
Coorte	Contiene l'anno di iscrizione dello studente
Data Nascita	Contiene la data di nascita dello studente
Genere	Contiene il genere dello studente o della studentessa
Diploma_Scuola_Superiore	Contiene il tipo di scuola scelta per gli studi superiori
voto_scuola_superiore	Contiene la valutazione ottenuta in sede di diploma di scuola superiore
area_geografica_scuolasuperiore	Contiene un valore indicante l'area geografica della scuola superiore. I valori possibili sono 5: Emilia-Romagna, Altre Regioni del Nord, Centro, Sud e Isole ed Estero.
area_geografica_residenza	Contiene un valore indicante l'area geografica di residenza. Sono possibili gli stessi valori del campo area_geografica_scuolasuperiore

Classe_ISEE	Contiene 9 valori possibili: non disponibile (Coorte 2016/17), ISEE non presentato, meno di 13.000, 13.000-23.000, 23.000-33.000, 33.000-45.000, 45.000-60.000, 60.000-70.000, oltre 70.000
Merito_ISEE	Contiene il valore di ISEE presentato durante la fase di iscrizione
OFA_assegnati	Contiene un valore booleano che descrive il fatto che siano stati assegnati o meno degli OFA (Obblighi Formativi Aggiuntivi)
OFA_superati	Contiene un valore booleano che descrive il fatto che siano stati superati o meno gli OFA
CdS	Descrive il corso di studi di appartenenza
TipoCorso	Descrive se si tratta di un corso di Laurea Triennale, Magistrale o Magistrale a ciclo unico
Ambito	Contiene un codice rappresentante la scuola di appartenenza del corso di studi (es. Scienze o Ingegeria)
Sede & Campus	Sono due colonne rappresentanti in testo e codice la sede nel quale il corso di studi viene erogato
Abbandoni	Contiene un valore booleano indicante se si è effettuato un abbandono del corso di studi
Passaggi	Contiene un valore booleano indicante se si è effettuato un abbandono del corso di studi
Trasferimenti	Contiene un valore booleano indicante se si è effettuato un abbandono del corso di studi

Un secondo file contiene invece per ogni studente la lista degli esami sostenuti. I campi contenuti sono:

Campo	**Descrizione**
CdS	Descrive il corso di studi di appartenenza
Cod.Materia	Codice identificativo della materia
Materia	Il nome della materia dell'esame sostenuto
Giorno Esame	Data in formato *gg/mm/yyyy* del giorno di sostenimento dell'esame
Voto_se_numerico	Campo rappresentante il voto ricevuto all'esame, se NaN si riferisce ad esami come le idoneità che non rilasciano voti.
CFUsuperati	Numero di CFU dell'esame sostenuto

1.3 Suddivisione del lavoro

Questo progetto è stato svolto in un gruppo di tre persone. Nella fase iniziale del lavoro abbiamo lavorato tutti e tre a stretto contatto per il preprocessing dei dati e la creazione di tutti i diversi dataset. Successivamente, ognuno ha applicato il proprio modello ai vari preprocessing precedentemente svolti e abbiamo poi unito e confrontato i risultati. Rispettivamente:

- Il modello di Regressione Logistica è stato utilizzato da Simone Boldrini
- Il modello Random Forest è stato utilizzato da Marco Benito Tomasone
- Il modello Support Vector Machine è stato utilizzato da Luca Genova

La motivazione della scelta di lavorare insieme per l'ottenimento dei vari preprocessamenti è legata alla volontà di unire le idee di tutti i componenti del gruppo e uniformarle per ottenere poi risultati comparabili tra i vari modelli.

1.4 Risultati Ottenuti

I risultati ottenuti dal nostro studio mostreranno, come ovviamente prevedibile, che è più facile prevedere l'abbandono di uno studente alla fine del primo anno piuttosto che a metà del primo anno, e che è più facile predire l'abbandono di uno studente a metà anno piuttosto che all'inizio dell'anno. Il modello che ottiene i risultati migliori (seppur con una differenza leggerissima rispetto agli altri) è invece il modello Random Forest.