Previsione di performance degli studenti con analisi dei dati dei registri elettronici

Giulio Angiani, Alberto Ferrari, Paolo Fornacciari, Monica Mordonini, and Michele Tomaiuolo

Dipartimento di Ingegneria e Architettura, Università di Parma
{giulio.angiani,alberto.ferrari,paolo.fornacciari,monica.mordonini,michele.tomaiuolo}@unipr.it

Abstract. Negli ultimi anni, la quantità di dati digitali presenti nelle scuole è aumentata in maniera esponenziale, anche come conseguenza dell'introduzione dei registri elettronici dove i docenti inseriscono i dati relativi alle attività degli studenti: assenze, valutazioni, tipologia delle prove assegnate. Tuttavia tali dati sono spesso presenti su diverse piattaforme e non è sempre facile utilizzarli per scopi scientifici. La nostra ricerca ha come focus principale proprio questo tipo di analisi. Dopo aver recuperato tali dati dalle scuole aderenti al progetto, abbiamo utilizzato alcune tecniche di data mining per analizzare comportamenti e performance degli studenti. I risultati indicano che: (i) è possibile anticipare la previsione di successo e di insuccesso già nei primi mesi dell'anno scolastico; e (ii) concentrandosi solo su un numero molto ristretto di discipline è anche possibile individuare con grande anticipo situazioni di disagio scolastico. In questa ottica stiamo realizzando un sistema web accessibile agli interessati che possa fornire una analisi delle situazioni degli studenti.

Keywords: Data mining; performance prediction; educational, classification.

1 Introduzione

Predire le performance degli studenti è una delle sfide più interessanti che una istituzione educativa può porsi oggi. Essere in grado di capire le difficoltà degli studenti prima possibile permette di mettere in atto le strategie didattiche migliori per prevenire un esito non positivo dell'anno scolastico. Le tecniche di Data Mining (DM) vengono applicate in molti campi di studio come finanza [10], sanità [16], previsioni meteo [15] e nella Social Network Analysis [3, 9] ma sono state recentemente applicate con ottimi risultati anche nell'ambito *educational*. L'applicazione dei DM in ambito educativo va solitamente sotto il nome di *Educational Data Mining (EDM)* [8]: questa è un'area di ricerca interdisciplinare ancora emergente che negli ultimi anni ha però ricevuto l'attenzione della comunità scientifica.

In one dei primi importanti lavori relativi all EDM [6], Baker individua cinque principali campi di studio: predizione, clusterizzazione, relazione, modellizzazione e estrazione di dati per la valutazione. Tuttavia, quasi tutti gli studi del campo EDM fanno riferimento al mondo accademico e solo in pochi casi alla fascia 14-18 anni [12].

In tutti i casi analizzati inoltre, i dati provengono da sondaggi effettuati fra gli studenti o contengono informazioni relative allo status sociale e famigliare degli studenti stessi.

Il lavoro mostrato in questo paper è la realizzazione di quanto annunciato in [2] durante l'edizione 2017 di Didamatica. In quella sede veniva presentata l'idea di fondo che giunge a primi effettivi risultati con questa pubblicazione.

La nostra ricerca si focalizza infatti esclusivamente sui dati dell'esperienza giornaliera di studenti e studentesse utilizzando dati contenuti nei registri elettronici di 10 scuole superiori italiane che hanno aderito a questo progetto di ricerca.

Nei registri elettronici sono contenuti dati relativi alle assenze giornaliere degli studenti, alle valutazioni nelle varie discipline, alle tipologie di prove sostenute.

Spesso questi dati sono sparsi fra molti sistemi informativi e gestiti da diversi provider ed è necessario aggregarli per poterli utilizzare in analisi scientifiche. In particolare noi abbiamo sviluppato un sistema web-based, chiamato ELDM (Electronic Logbook Data Mining) che permette agli utenti autorizzati di (i) condividere i propri dati, opportunamente deidentificati e anonimizzati, con la piattaforma e (ii) di ottenere alcune analisi dettagliate relative alle situazioni presenti nelle singole scuole.

Il resto dell'articolo è strutturato come segue: la sezione 2 si occupa di esporre una breve review della letteratura in essere; la sezione 3 descrive le modalità di raccolta dei dati e le metodologie di analisi; risultati e discussione sono presentanti nella sezione 4; infine la sezione 5 fornisce alcune osservazioni sul progetto e su eventuali sviluppi futuri.

2 Letteratura in merito

Uno dei primi studi applicati all'EDM è da attribuire a Kapur[11] et al. Nel loro studio sono stati applicati vari algoritmi a dati relativi a 480 studenti iscritti all'università dell'India. Per ogni studente sono state individuate 16 caratteristiche, parte di esse relative a fattori familiari. La ricerca ha mostrato come l'algoritmo J48 e Random Forest siano i migliori per individuare una corretta previsione di rendimento. In [12], sono invece stati effettuati esperimenti di previsione di abbandono per 419 studenti di una scuola superiore del Messico. Anche in questo caso vengono utilizzati dati relativi alla condizione sociale degli studenti. A causa del basso numero di casi analizzati il data-set è ampiamente sbilanciato. In [5], Asif *et al.* analizzano invece studi universitari. Questo paper dimostra come, concentrandosi su un numero limitato di corsi molto significativi, sia possibile evidenziare in anticipo difficoltà o problemi a supporto degli studenti. Alcuni studi mettono in relazione la capacità di spesa famigliare con i

risultati scolastici [7] mentre in altri viene mostrato come utilizzare tecniche di *ensemble learning* e di *machine learning* allo specifico dominio utilizzando dati di corsi di università americane come UCLA e altre [17, 4]. In [13] sono indicate delle analisi comparative delle performance di studenti utilizzando alberi di decisione mentre in [14] Prasada Rao *et al.* mettono a confronto algoritmi come J48, Naïve Bayes e Random forest. In questo caso il dataset consiste in soli 200 casi di studenti universitari di scienze e ingegneria. Random Forest dimostra di essere il miglior algoritmo per la predizione di performance in questo caso.

Il nostro lavoro conferma questi risultati ma con una decisiva differenza: i dati in nostro possesso non sono in alcun modo relativi alla condizione socio-economica degli studenti e delle studentesse né ad altre informazioni extra-scolastiche ma solo ed esclusivamente connessi alla loro attività didattica grazie all'analisi di valutazioni e assenze inserite giorno per giorno all'interno dei registri elettronici dai docenti. Nonostante questo i risultati di predizione delle performance superano l'80% di accuratezza già con dati dei primi mesi dell'anno scolastico e permettono di segnalare situazioni di disagio per poter attuare politiche didattiche e di recupero prima possibile.

3 Dati e metodologia

I dati utilizzati per la nostra ricerca sono stati estratti dai registri elettronici di 10 scuole superiori, situate in diverse parti d'Italia. Tutti i dati sono stati resi anonimi in conformità con le attuali leggi sulla privacy in Italia [1]. Tutte le informazioni sono relative ai voti ottenuti dagli studenti nelle verifiche e alla loro frequenza alle lezioni. Sono inoltre presenti dati relativi agli esiti di fine anno e alle valutazioni di matematica e italiano di fine periodo. Le informazioni su voti e presenze sono state utilizzate per addestrare alcuni classificatori con lo scopo di prevedere l'esito finale.

Preparazione dei dati. Dai dati presenti sono state utilizzate solo le valutazioni in scala 1 a 10 con anche valori decimali. Valori non appartenenti all'interavallo [1-10] sono stati scartati.

Le materie sono state suddivise in 6 gruppi:

1. Italiano (ita)
2. Matematica (mat)
3. Inglese (eng)
4. Storia (his)
5. Materie di indirizzo (cou)
6. Altre discipline (oth)

Il raggruppamento delle materie è stato necessario per analizzare gli studenti appartenenti a corsi di studio non omogenei. I primi quattro gruppi sono comuni a tutti i corsi delle scuole superiori italiane, ma ogni corso ha anche le sue materie specifiche: queste sono state raggruppate nel quinto gruppo (*cou*). Il sesto gruppo (*oth*) contiene tutti le discipline rimanenti. [1]

[1] E.g.: 'Informatica' appartiene al 5° gruppo per gli studenti di un tecnico informatico, ma al 6° gruppo per gli studenti di un liceo linguistico

Selezione e trasformazione delle features. A partire dai dati grezzi giornalieri, abbiamo costruito le caratteristiche degli studenti, raggruppando i dati per ogni mese con il seguente metodo: per ogni gruppo, abbiamo calcolato il punteggio medio, raccolto nel periodo dal 15 settembre (inizio della scuola) fino alla fine di ogni mese, da Ottobre a Maggio. Il nome assegnato a queste features ha il seguente formato per i valori relativo a un singolo studente: *<materia>_<mese>*[2] e il seguente formato, per la media ottenuta utilizzando i voti di tutti i suoi compagni di classe per lo stesso periodo e per lo stesso argomento: *<materia>_<mese>_grp*[3].

La tabella 1 mostra la lista *F1* di tutte le caratteristiche ottenute con questa metodologia. F1 contiene 6*8*2 = 96 features.

Il secondo gruppo di caratteristiche contiene informazioni sulla frequenza scolastica di uno studente e sulla frequenza media dei suoi compagni di classe. Ogni caratteristica mostra la frequenza scolastica di uno studente in un certo periodo, in numero di giorni. Come nel caso precedente, queste features hanno il seguente formato: *abs_<mese>*; and *abs_avg_<mese>_grp*.

Il terzo set di funzionalità calcolate è correlato all'andamento di uno studente in un periodo. Per indicare il valore del trend, abbiamo utilizzato i coefficienti m, c e dev di una retta di regressione lineare calcolata usando le valutazioni di uno studente in quel periodo[4]. Anche in questo caso le stesse caratteristiche sono state calcolate anche per ciascun gruppo di studenti (288 features).

L'ultima serie di funzionalità contiene solo dati sulla scuola, l'anno, l'anno scolastico, il corso di studi e alcuni dati sui voti di fine anno degli studenti (in totale 10 valori). L'insieme delle caratteristiche utilizzate contiene quindi al massimo 410 elementi per ogni studente.

La tabella 2 mostra la lista *F2* di tutte le features associate alle assenze. F2 contiene 8 * 2 = 16 features.

Un esempio di calcolo della retta di regressione a partire dai voti ottenuti nella materia "Lingua Inglese" è mostrata in Figura 1.

I valori usati per tale calcolo sono mostrati in Tabella 3.

Dataset finale. L'intero dataset contiene 13151 differenti istanze relative a 10342 diversi studenti di 10 scuole superiori italiane. Esistono più istanze che studenti in quanto alcune scuole hanno fornito dati di più anni scolastici.

Ogni istanza è formata al massimo da 410 valori reali, uno per ogni feature dell'insieme F sopra indicato. Ogni istanza è anche associata all'esito di giugno, obbligatorio nel caso di esperimenti di previsione della performance di fine anno. Il valore dell'esito finale può essere POSITIVO, NEGATIVO o SOSPESO nel caso di studenti dei primi 4 anni, POSITIVO o NEGATIVO per studenti dell'ultimo anno di corso.

[2] *ita_nov* è la feature per il calcolo della media in italiano a fine novembre

[3] *ita_nov_grp* è la feature per il calcolo della media in italiano a fine novembre di tutta la classe

[4] *m* è il coefficiente angolare, *c* il valore di intersezione con l'asse delle y, *dev* è la deviazione standard della retta di regressione lineare

Table 1. Lista delle 96 features ottenute dalle medie dei voti degli studenti e dei gruppi di appartenenza fra Ottobre e Maggio nei sei gruppi di materie individuati.

	Materia	Mese	Feature	Feature gruppo
1	Italiano	Ottobre	ita_oct	ita_oct_grp
2	Italiano	Novembre	ita_nov	ita_nov_grp
3	Italiano	Dicembre	ita_dec	ita_dec_grp
...				
8	Italiano	Maggio	ita_may	ita_may_grp
9	Matematica	Ottobre	mat_oct	mat_oct_grp
...				
16	Matematica	Maggio	mat_may	mat_may_grp
17	Inglese	Ottobre	eng_oct	eng_oct_grp
...				
...				
48	Altro	Maggio	oth_may	oth_may_grp

Table 2. Lista delle 16 features costruite con le informazioni delle assenze fra Ottobre e Maggio.

	Mese	Feature	Feature gruppo
1	Ottobre	abs_oct	abs_avg_oct_grp
2	Novembre	abs_nov	abs_avg_nov_grp
...			
8	Maggio	abs_may	abs_avg_may_grp

Nella tabella 4 è indicata la distribuzione degli studenti in funzione dell'esito di fine anno.

Avendo a disposizione un numero di istanze significativo per ogni tipologia (il minimo valore è di 1100 per gli studenti con esito negativo) è stato possibile ottenere sempre un dataset bilanciato (ovvero con un numero di istanze uguale per ogni tipologia) in ogni esperimento.

Sono stati eliminati dal dataset tutte le istanze con un numero di valori non significativi superiore a 20 (su 410 totali).

In questo modo siamo riusciti ad ottenere più di 1000 elementi validi per ogni classe di appartenenza (POSITIVO, NEGATIVO o SOSPESO) e quindi più di 3000 complessivamente. Di questi 1500 sono stati utilizzati nell'addestramento del sistema (training-set) e altrettanti per il test (test-set)

4 Risultati e discussione

Nel nostro lavoro abbiamo effettuato vari esperimenti: nel primo ci siamo concentrati sulla previsione dell'esito finale.

Confronto fra differenti metodi di classificazione. Per questo obiettivo abbiamo usato tre differenti tecniche di classificazione: abbiamo confrontato i risultati usando quattro differenti algoritmi di classificazione: Random Forest,

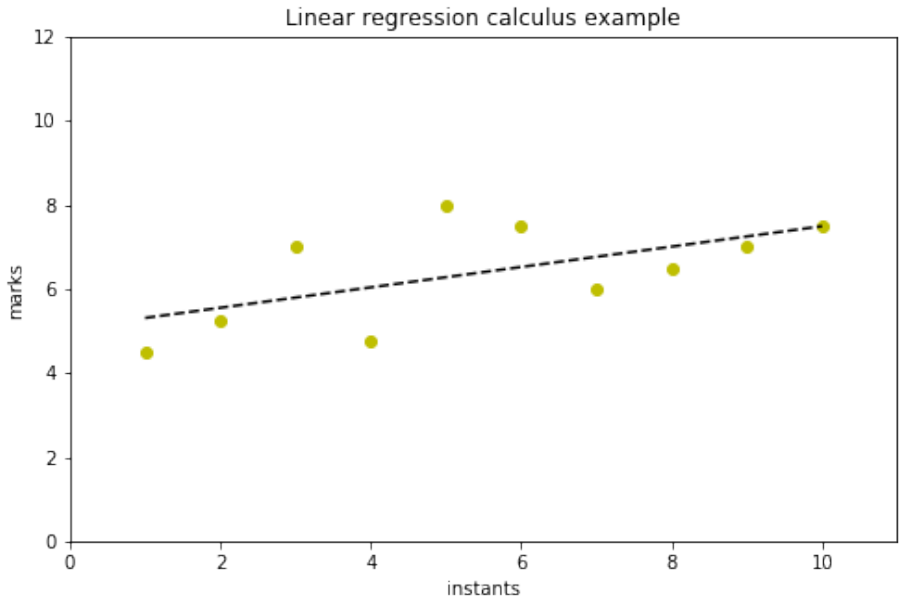

Fig. 1. Calcolo retta di regressione.

Neural Network, SVM e KNN. L'esperimento è stato effettuato considerando due periodi dell'anno scolastico. Il primo con i dati da ottobre a dicembre, il secondo con quelli da gennaio a marzo.

In entrambi i casi sono state utilizzate 1000 istanze per ogni categoria (POSITIVO, NEGATIVO, SOSPESO). 500 di queste sono state usate per il training e 500 per il test set.

Le features estratte sono state 151 avendo ristretto l'analisi ad un intervallo di 3 mesi.

Le immagini 2 e 3 indicano i risultati massimi di accuratezza ottenuti utilizzando i 4 classificatori.

In entrambi i casi i risultati migliori in termine di accuratezza sono stati raggiunti con Random Forest che arriva quasi al 90% di precisione.

Scelta delle features più significative Altro risultato molto interessante che abbiamo raggiunto è stato quello di capire quali sono le feature più significative per l'individuazione dell'esito finale e quindi per segnalare prima possibile eventuali situazioni di difficoltà. Sempre dal grafico 2 si può osservare come il massimo valore di accuratezza si raggiunge con sole 60 caratteristiche (delle 151 calcolate).

Predizione per l'esito dell'ultimo anno Il secondo esperimento ha preso invece in esame i dati del solo ultimo anno di corso. In questo caso l'accuratezza della previsione a dicembre è stata di circa il 95%.

Table 3. Dati usati per il calcolo retta di regressione in esempio.

Data	Voto
2015-10-19	4.5
2015-11-17	5.25
2015-11-30	7
2016-01-15	4.75
2016-01-26	8
2016-02-23	7.5
2016-03-04	6
2016-03-24	6.5
2016-04-13	7
2016-05-05	7.5

Table 4. Distribuzione delle istanze secondo l'esito finale.

Esito finale	Numero di istanze	Percentuale
POSITIVO	10609	80,67 %
NEGATIVO	1100	8,36 %
SOSPESO	1442	10,96 %

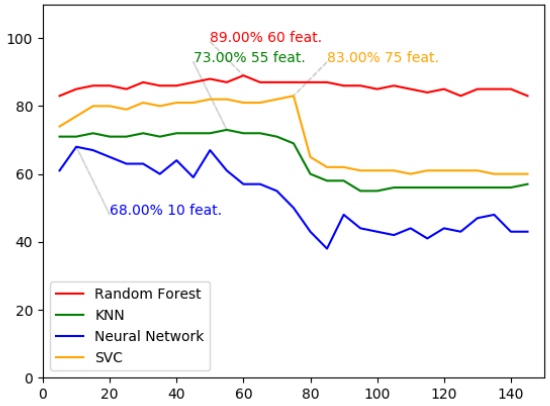

Fig. 2. Accuratezza con numero di features diverse e classificatore per il periodo ottobre-dicembre - Massimo raggiunto con 60 features.

Anche in questo caso i migliori risultati sono stati raggiunti utilizzando Random Forest in accordo con altri lavori citati [14, 11].

Relazione fra materie e previsione Un altro interessante risultato è stato osservare quali siano le caratteristiche che più influenzano il risultato finale. In

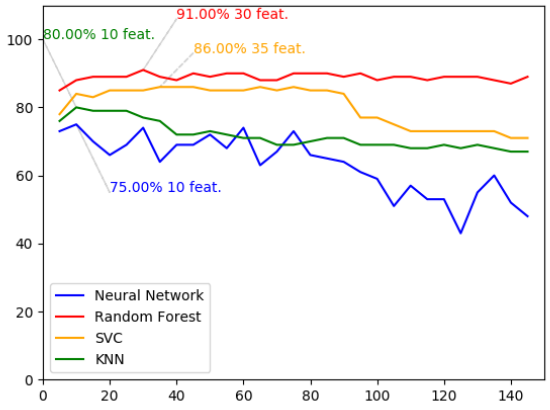

Fig. 3. Accuratezza con numero di features diverse e classificatore per il periodo gennaio-marzo - Massimo raggiunto con 30 features.

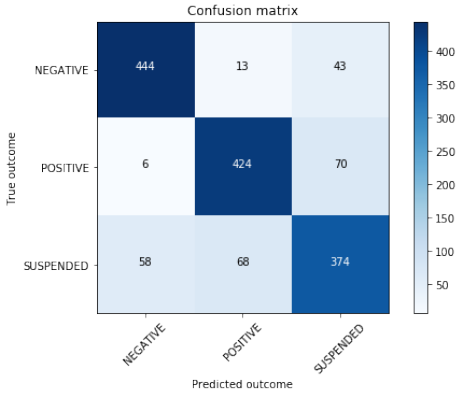

Fig. 4. Matrice di confusione su test set con algoritmo RF e con le 30 features più significative - periodo ottobre/dicembre - accuratezza 82%.

Tab. 5 sono indicate le 10 features più significative per l'analisi di predizione effettuata.

Questi indicatori permettono di evidenziare subito situazioni di difficoltà analizzando pochi dati già nei primi mesi, considerando anche che l'accuratezza cambia molto poco utilizzando 10 features piuttosto che 60 che rappresenta il massimo per il periodo ottobre-dicembre (fig. 2)

Table 5. Le 10 features più significative ordinate secondo l'algoritmo Analysis of Variance (ANOVA) ed il punteggio relativo.

Rank	Value	Feature
1	0.08129	ind_nov
2	0.05663	ind_dec
3	0.03855	trend_c_ind_dec
4	0.03289	trend_c_ind_oct
5	0.03216	sto_dec
6	0.02964	mat_dec
7	0.02576	trend_c_mat_dec
8	0.02559	ita_nov
9	0.0221	eng_nov
10	0.02144	ind_oct

5 Conclusioni

Sulla base dell'esperienza personale, docenti e dirigenti tendono a farsi una idea di successo degli studenti già nei primi mesi di scuola. Il nostro lavoro, analizzando dati reali della vita scolastica dei ragazzi, propone uno strumento validato e obiettivo per evidenziare diversamente eventuali problematicità già nel primo periodo di attività.

E' possibile infatti far emergere alcuni elementi altamente significativi per segnalare difficoltà di apprendimento, laddove non fosse già evidente per i docenti, al fine di permettere un intervento tempestivo di recupero e facilitazione.

Tale possibilità potrebbe essere utilizzata anche per suggerire passaggi da una scuola ad un'altra e per attivare soluzioni di contrasto alla dispersione scolastica.

Ringraziamenti

Questa ricerca è stata patrocinata dall'Assemblea Legislativa dell'Emilia-Romagna. Gli autori inoltre ringraziano il Gruppo Spaggiari e la Argo Software per il supporto tecnico necessario a sviluppare moduli software a beneficio delle scuole per l'esportazione dei dati presenti nei registri elettronici scolastci. Uno speciale ringraziamento alle scuole che hanno aderito al progetto di ricerca.

Riferimenti bibliografici

1. Codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e scientifici - deontological and good practice code for treating personal data used in scientifical and statistical issues. Gazzetta Ufficiale della Repubblica Italiana (190) (2004).
2. Angiani, G., Ferrari, A., Giannotti, F., Poggi, A., Salvatori, E.: Electronic logbook data mining. In: Didamatica 2017. vol. 1, pp. 1–4. AICA (2017).

3. Angiani, G., Fornacciari, P., Iotti, E., Mordonini, M., Tomaiuolo, M.: Models of Participation in Social Networks., pp. 196–224. edited by Michael A. Brown Sr., IGI Global (2017).
4. Asif, Hina, Haque: Predicting student academic performance using data mining methods. International Journal of Computer Science and Network Security 17(5), 187–191 (2017).
5. Asif, Merceron, Ali, Haider: Analyzing undergraduate students' performance using educational data mining. Computers & Education 113(3), 177–194 (2017).
6. Baker, et al.: Data mining for education. International encyclopedia of education, Oxford, UK: Elsevier, 7(190), 112–118 (2010).
7. Daud, Aljohani, Abbasi, Lytras, Abbas, Alowibdi: Predicting student performance using advanced learning analytics. International World Wide Web Conference Committee (IW3C2) pp. 415–421 (2017).
8. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5-6), 304–317 (2012).
9. Fornacciari, P., Mordonini, M. and Tomauiolo, M. Social network and sentiment analysis on Twitter: Towards a combined approach. CEUR Workshop Proceedings Volume 1489 (KDWeb2015), pp 53-64 (2015).
10. Jadhav, S., He, H., Jenkins, K.W.: An academic review: Applications of data mining techniques in finance industry. International Journal of Soft Computing and Artificial Intelligence 4(1), 79–95 (2017).
11. Kapur, Ahluwalia, Sathyaraj: Comparative study on marks prediction using data mining and classification algorithms. International Journal of Advanced Research in Computer Science 8(3), 632–636 (2017).
12. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H., Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Systems 33(1), 107–124 (2016).
13. Pereira, Pai, Fernandes: A comparative analysis of decision tree algorithms for predicting student's performance. International Journal of Engineering Science and Computing 7(4), 10489–10492 (2017).
14. Rao, P., Sekhara, B.Ramesh, et al.: Predicting learning behavior of students using classification techniques. International Journal of Computer Applications 7(139), 112–118 (2016).
15. Venkatesh, S., Chandrakala, D.: A survey on predictive analysis of weather forecast. weather (2017).
16. Vidhu, R., Kiruthika, S.: A survey on data mining techniques and their comparison approaches for healthcare. Data Mining and Knowledge Engineering 9(1), 14–19 (2017).
17. Xu, Han, Marcu, van der Schaar: Progressive prediction of student performance in college programs. In: Thirty-First AAAI Conference on Artificial Intelligence. pp. 1604–1610 (2017).