Applicazione del machine learning ai learning analytics della piattaforma Moodle per creare gruppi eterogenei nei corsi on-line

Application of machine learning to the learning analytics of the Moodle platform to create heterogeneous groups in on-line courses

Giacomo Nalli
School of Science and Technology, Computer Science section, University of Camerino
giacomo.nalli@unicam.it
Leonardo Mostarda
School of Science and Technology, Computer Science section, University of Camerino
leonardo.mostarda@unicam.it
Andrea Perali
School of Pharmaceutical Sciences and Health Production, University of Camerino
andrea.perali@unicam.it
Sebastiano Pilati
School of Science and Technology, Physics Section, University of Camerino
sebastiano.pilati@unicam.it
Daniela Amendola
School of Biosciences and Veterinary Medicine, University of Camerino
daniela.amendola@unicam.it

In university courses to promote collaborative activities among students, on-line learning environments such as e-learning platforms are used. Effective collaborative activities involve the creation of heterogeneous groups of 4 or 5 students. In the university context the formation of groups is difficult due to the high number of students. Groups are often unbalanced and not very functional if chosen randomly. Some e-learning platforms, such as Moodle, lack an intelligent mechanism that allows the automatic creation of heterogeneous groups of students. We applied clustering algorithms on Moodle learning analytics (LA) that allowed to build groupings that identify the different characteristics of students based on their behaviors kept on the platform. Therefore we have developed an intelligent numerical tool which, using clusters obtained from Machine Learning on the LA, generates heterogeneous groups. These groups are made available on the platform for the teacher. The project will conclude with the development of a Moodle plugin to automate the exchange of data and information between the Machine Learning algorithm and the Moodle platform.

Keywords: Learning Analytics; Machine learning; Moodle; Clustering; Gruppi

Nei percorsi universitari, per favorire le attività collaborative tra gli studenti, vengono utilizzati ambienti di apprendimento on-line come le piattaforme e-learning. Attività collaborative efficaci prevedono la creazione di gruppi eterogenei di 4 o 5 studenti. Nel contesto universitario la formazione dei gruppi è difficile per l'elevato numero di studenti. Se scelti in maniera casuale, spesso i gruppi risultano sbilanciati e poco funzionali. Alcune piattaforme e-learning, ad esempio Moodle, mancano di un meccanismo "intelligente" che permetta di creare in automatico gruppi eterogenei di studenti. Il nostro lavoro consiste nel realizzare un software in Python in grado di creare gruppi eterogenei di studenti, utilizzando tecniche di Machine Learning con i dati estratti da Moodle. Abbiamo applicato

 – ISSN 2038-9744 (on line) – DOI 10.7346/SIRD-2S2019-P158

algoritmi di clustering sui learning analytics (LA) di Moodle che hanno permesso di costruire dei raggruppamenti che identificano le caratteristiche degli studenti in base ai loro comportamenti in piattaforma. Abbiamo quindi sviluppato uno strumento numerico "intelligente" che, utilizzando i cluster ottenuti dal Machine Learning sui LA, genera gruppi eterogenei. Questi gruppi vengono messi a disposizione in piattaforma per il docente. Il progetto si concluderà con lo sviluppo di un plugin di Moodle per automatizzare lo scambio di dati e informazioni tra l'algoritmo di Machine Learning e la piattaforma Moodle.

Parole chiave: Learning Analytic; Machine learning; Moodle; Clustering; groups

1. Introduzione

Le attività collaborative sono una metodologia didattica estremamente significativa per gli studenti poiché, promuovendo un apprendimento attivo (Foote, 2009), migliorano i risultati dell'apprendimento oltre che sviluppare le loro abilità sociali come il processo decisionale, la comunicazione, le capacità collaborative ed il pensiero critico (Smith and MacGregor, 2009; Stevens, Levy, 2005; Amendola, Miceli, 2018). A causa del gran numero di studenti iscritti nei percorsi universitari, talvolta si possono riscontrare difficoltà a livello organizzativo nella progettazione di attività collaborative. Oggi grazie allo sviluppo delle tecnologie digitali è possibile organizzare esperienze di apprendimento collaborativo on-line in maniera più flessibile, sia per gli studenti che per i docenti (Abedin et al., 2012). Nei contesti formativi universitari, per favorire le attività collaborative tra gli studenti, vengono spesso utilizzati ambienti di apprendimento on-line come le piattaforme e-learning che, come ambienti web 2.0, enfatizzano la partecipazione, la connessione, la condivisione di conoscenze ed idee tra gli studenti grazie a strumenti appositamente progettati (McLough, 2007).

La piattaforma e-learning Moodle (Modular Object-Oriented Dynamic Learning Environment, ambiente per l'apprendimento modulare, dinamico, orientato ad oggetti), molto utilizzata dalle università italiane e straniere, offre diversi strumenti per stimolare la partecipazione, l'interazione, la negoziazione e la collaborazione tra studenti, come ad esempio il forum, il wiki ed il workshop. Tali strumenti permettono al docente di poter effettuare e gestire le attività in maniera semplice e automatizzata.

Sia in ambiente on-line che in presenza, uno dei fattori fondamentali che può influenzare il successo dell'apprendimento collaborativo è la composizione di gruppi in base al numero ed all'eterogeneità dei componenti. In riferimento al numero dei componenti, gruppi accettabili vengono considerati quelli formati da almeno 3 o più persone, anche se in generale gruppi da 4 o 5 membri risultano essere più efficaci (Burke, 2011).

I gruppi possono essere formati anche da due persone, ma questo numero non incoraggia il lavoro di gruppo perché non è sufficiente per generare creatività e varietà di idee (Csernica et al., 2002). Un'altra caratteristica fondamentale per il successo nei lavori di gruppo è sicuramente l'eterogeneità degli studenti in termini di risorse cognitive, caratteristiche e comportamenti (Nijstad, De Dreu, 2002).

Tuttavia formare gruppi ottimali di studenti per attività collaborative non è semplice. Normalmente si utilizzano diversi approcci che però non garantiscono sempre la formazione di gruppi eterogenei: selezione casuale, selezione automatica e selezione del docente (Sadeghi, Kardan, 2016). Nel primo caso il docente assegna gli studenti casualmente ai gruppi, manualmente o tramite un sistema informatico. Questo è il metodo più semplice e veloce che consente di mescolare tutti gli studenti con la speranza di raggiungere l'eterogeneità all'interno dei gruppi (Bacon et al., 2001), fondamentale per il miglioramento delle performance sociali e cognitive, soprattutto per studenti meno motivati.

La formazione del gruppo può essere condotta anche attraverso l'approccio di auto-selezione da parte degli studenti che consente di avere gruppi con un'elevata empatia, ma non sempre eterogenei sotto il profilo pedagogico.

Il terzo approccio, che potrebbe garantire la formazione di gruppi eterogenei, è la selezione da parte del docente sulla base di caratteristiche prestabilite, come per esempio conoscenza, abilità, interessi e stile di apprendimento (Jackson et al., 1995).

Quest'ultimo metodo, che risulta essere il più funzionale per la realizzazione di gruppi eterogenei, in ambiente universitario è di difficile attuazione. Per il docente universitario, infatti, l'identificazione degli studenti che frequentano l'aula, in base a determinate caratteristiche e comportamenti, risulta essere complicata non solo per l'elevato numero di partecipanti ma anche in quanto i corsi hanno una durata relativamente breve e non sempre hanno frequenza obbligatoria.

Nei percorsi universitari, dove per le attività didattiche vengono utilizzate anche le piattaforme e-learning, l'elaborazione dei Learning Analytics (LA) prodotti può essere utile per creare gruppi eterogenei di studenti sulla base delle loro caratteristiche e comportamenti in piattaforma.

Il punto di debolezza è che in questi ambienti on-line manca ancora un meccanismo intelligente che permetta di creare questi gruppi in modo automatico, facilitando in tal modo il lavoro del docente.

Negli ultimi anni alcuni gruppi di ricerca hanno iniziato a lavorare su progetti di Machine Learning applicati ai LA estratti dalle piattaforme e-learning per creare gruppi eterogenei in modo automatico: al-

cuni autori si sono basati sui dati estratti da un questionario on-line (Razmerita, Brun, 2011), altri sui dati estratti dall'interazione degli studenti all'interno dello strumento forum di Moodle (Maina et al., 2017). I lavori presenti in letteratura per la formazione di gruppi eterogenei on-line si basano principalmente sull'elaborazione di LA estratti da un'unica attività svolta dagli studenti nella piattaforma e-learning.

Dutt et al. (2017) nel loro lavoro passano in rassegna la letteratura esistente sul clustering dei dati dai processi di apprendimento. In questa rassegna la possibilità di realizzare una classificazione di studenti in base ai loro stili di apprendimento viene presentata come applicazione di successo delle tecniche di clustering dei dati provenienti dai processi di apprendimento. Un'importante conclusione ribadita da Dutt et al. è che questi dati sono spesso non-indipendenti e con una struttura gerarchica multi-livello e quindi la scelta delle variabili per la separazione dei cluster assume un ruolo chiave e deve essere effettuata dal ricercatore in modo molto scrupoloso per ottenere risultati coerenti. Una sezione della rassegna di Dutt et al. è dedicata alle tecniche di clustering applicate ai dati provenienti da corsi erogati in e-learning, com'è il caso del nostro lavoro. Si evidenzia che numerosi articoli sono presenti in letteratura su questo argomento in quanto i corsi e-learning forniscono molti dati da poter utilizzare grazie agli strumenti di tracciamento delle piattaforme e-learning. Interessante sottolineare applicazioni delle suddette tecniche allo studio del problem-solving per individuare in modo automatico gli stili di risoluzione e di apprendimento degli studenti. La sezione di questa rassegna dedicata al collaborative learning evidenzia l'efficacia dell'algoritmo K-means per la formazione di gruppi e come gli ambienti e-learning siano i più adatti per gli studenti che collaborano attivamente. Conclusioni queste che sono in linea con i metodi adottati e i risultati ottenuti nel nostro lavoro.

Infine, un importante confronto tra i LA ed il cosiddetto Educational Data Mining (EDM) viene discusso da Papamitsiou & Economides (2014). Per quanto riguarda l'approccio per acquisire informazioni sui processi di apprendimento, LA adotta un quadro olistico, cercando di comprendere i sistemi nella loro piena complessità. D'altra parte, EDM adotta un punto di vista riduzionista analizzando i singoli componenti del processo di apprendimento, cercando nuovi modelli nei dati e modifica dei rispettivi algoritmi. Queste due ricerche sono complementari ed al fine di catturare l'intero quadro dei processi di apprendimento degli studenti, entrambi gli approcci andrebbero perseguiti.

Il nostro progetto di ricerca consiste nell'elaborare insieme i LA estratti dalle molteplici attività che gli studenti svolgono durante un percorso on-line, sfruttando al meglio le correlazioni presenti nei dati.

Lo scopo è quello di ottenere una visione complessiva dei comportamenti degli studenti per meglio caratterizzarli, permettendoci così di creare dei gruppi quanto più possibili eterogenei. Il nostro approccio consiste nella creazione di un'applicazione informatica che permetta la realizzazione di gruppi eterogenei in modo automatico, usando tecniche di Machine Learning non supervisionato (Kotsiantis, 2007) applicate ai LA prodotti dagli studenti durante la fruizione di un percorso on-line in Moodle ed elaborati tramite il linguaggio di programmazione Python. La scelta iniziale dei LA da utilizzare viene fatta dal Docente, tramite caselle di spunta visualizzate dalla piattaforma e-learning Moodle, nel primo step del processo. Lo scopo è di poter selezionare da subito i dati che corrispondono ai criteri generali in base ai quali il Docente intende comporre i gruppi. La scelta dei LA fatta dal Docente includerà in tal modo la sua idea didattica e costituirà il set iniziale sul quale andrà ad operare il software di clusterizzazione automatica realizzato e descritto in questo articolo. Al Docente rimarrà pertanto un controllo a monte del processo di formazione automatica dei gruppi, con possibilità di variare le sue scelte iniziali dalle quali far partire, o ripartire, il software e di analizzarne le relazioni di causa-effetto così ottenute, in primis la stabilità dei risultati, ovvero la composizione dei gruppi, al variare delle condizioni iniziali.

In questo articolo, come corso pilota per il nostro progetto di ricerca è stato scelto un percorso laboratoriale on-line di fisica per il primo anno del corso di laurea in lingua inglese in Biosciences and Biotecnology, organizzato in una parte individuale ed una parte collaborativa che richiede la formazione di gruppi di lavoro.

Il nostro progetto di ricerca si sviluppa in due fasi:

1. applicazione di algoritmi di clustering ai Learning Analytics di Moodle estratti alla fine del percorso didattico individuale per la creazione di raggruppamenti omogenei al loro interno basati sulle caratteristiche e comportamenti in piattaforma simili degli studenti. Per verificare l'efficienza delle tecniche di clustering per la formazione dei raggruppamenti omogenei di studenti, sono stati messi a confronto le caratteristiche dei diversi raggruppamenti con gli esiti di un elaborato svolto alla fine del percorso individuale;

2. sviluppo di un software intelligente che, utilizzando i raggruppamenti ottenuti, distribuisce automaticamente gli studenti degli stessi raggruppamenti in diversi gruppi che risultano in questo modo eterogenei al loro interno. Dopo la formazione dei gruppi eterogenei il software comunica al docente automaticamente i gruppi creati.

2. Aspetti Metodologici

2.1. *Descrizione dell'attività formativa*

Il laboratorio online di fisica è stato erogato tramite la piattaforma e-learning Moodle di Ateneo e vi hanno partecipato 55 studenti internazionali (19 maschi e 36 femmine).

Il percorso didattico è strutturato in due parti. Una prima parte individuale, della durata di un mese, caratterizzata da 5 video esperimenti, realizzati nei nostri laboratori, relativi alla forza elastica e all'oscillatore armonico. Ogni video esperimento dura in media 10 minuti per un totale di 50 minuti.

I video esperimenti on-line consentono agli studenti di esaminare i fenomeni elastici, raccogliendo ed elaborando dati sperimentali utilizzando il software gratuito Gnuplot, al fine di ottenere un fit dei dati e trovare la formulazione matematica della legge fisica corrispondente. Per l'utilizzo del software sono disponibili in piattaforma due video tutorial: i) come installare Gnuplot; ii) come utilizzare Gnuplot per creare un grafico, preparare figure ed elaborare i dati. I video tutorial durano in totale 15 minuti. Sono presenti, inoltre, altri files (pdf e pagine web) con testo ed immagini, che contengono ulteriori spiegazioni relative a Gnuplot ed alla parte teorica dei video esperimenti.

La parte individuale si conclude con la realizzazione di un elaborato finale, da caricare in piattaforma attraverso il modulo consegna di Moodle, dove lo studente deve rispondere ad un set di domande e risolvere esercizi guidati dai video tutorial e dai video esperimenti. L'elaborato viene poi valutato dal docente.

La seconda parte, caratterizzata dall'attività collaborativa, consiste nella realizzazione di un report finale sugli argomenti proposti nella prima parte del percorso, da realizzare in gruppo. Per rendere l'attività collaborativa più efficace in questa fase del percorso vengono creati gruppi eterogenei formati da 4 o 5 studenti ognuno, come illustrato di seguito.

2.2. *Machine Learning*

Per la realizzazione dei gruppi abbiamo applicato tecniche di Machine Learning non supervisionato ai Learning Analytics prodotti dagli studenti in piattaforma durante lo svolgimento del percorso individuale. Il Machine Learning è un insieme di tecniche sviluppate nel campo dell'intelligenza artificiale che include al suo interno vari modelli sta-

tistici complessi e i corrispondenti metodi di ottimizzazione. L'obiettivo è quello di costruire algoritmi che possano estrarre informazione utile da grosse moli di dati a disposizione ed individuare delle correlazioni tra di essi, fornendo all'utilizzatore un modello in grado di effettuare predizioni accurate su contesti nuovi. Spesso, ma non sempre, questi modelli sono costruiti tramite reti neurali artificiali.

Nell'ambito della didattica on-line, tecniche di Machine Learning possono essere usate per diversi scopi, come per esempio informare i docenti sull'andamento del corso (Feng, Heffernan, 2005) o nella predizione del voto finale dello studente (Lopèz et al., 2012).

Tra le tecniche di Machine Learning, definite di apprendimento non supervisionato, troviamo in particolare gli algoritmi di clustering.

Questa tecnica permette di individuare un gruppo di oggetti che hanno caratteristiche simili, secondo criteri definiti a priori, e di assegnarli ad un cluster o raggruppamento.

Negli ambienti di apprendimento on-line, il clustering può essere usato per trovare raggruppamenti di studenti con simili caratteristiche, relative, ad esempio, ai livelli di apprendimento (Bovo et al., 2013).

Sono disponibili diversi algoritmi di clustering che consentono di formare gruppi di studenti sulla base del loro comportamento durante la fruizione di un corso online in piattaforma e-learning. Questi algoritmi risultano essere utilizzati per diverse finalità: per esempio l'algoritmo "Self Organizing Maps" permette di raggruppare gli studenti in base al loro background, oppure l'algoritmo "Fuzzy c-means" che permette di raggruppare studenti in base alla loro personalità e strategia di apprendimento (Vellido et al., 2010). La nostra scelta è ricaduta sull'algoritmo di clustering K-Means perché oltre ad essere un algoritmo molto utilizzato e di facile uso grazie alla libreria scikit learn che ne fornisce l'implementazione, è un algoritmo basato sull' utilizzo della distanza euclidea per la computazione che assegna un determinato elemento (in questo caso lo studente) a uno specifico cluster. Queste tipologie di algoritmo risultano essere ottimali per il raggruppamento degli studenti in base al comportamento relativo alla navigazione online effettuata tra le varie attività all'interno di un corso e-learning (Vellido et al.) e per organizzare in cluster i differenti comportamenti degli studenti tenuti nella piattaforma e-learning (Dutt et al., 2017).

Nel nostro lavoro per la realizzazione dei gruppi eterogenei vengono applicate prima tecniche di clustering, utilizzando l'algoritmo K-means, per organizzare dei raggruppamenti di studenti con caratteristiche simili (comportamento durante la fruizione del percorso on-line) ricavate dall'elaborazione dei LA forniti dalla piattaforma Moodle.

Successivamente, un software intelligente, distribuisce automaticamente gli studenti degli stessi raggruppamenti in diversi gruppi che risultano in questo modo eterogenei al loro interno.

3. Formazione di gruppi eterogenei

In questa sezione descriviamo la metodologia utilizzata per la raccolta e l'elaborazione dei dati sul comportamento degli studenti estratti dalla piattaforma e-learning al fine di costituire gruppi di studenti al loro interno eterogenei per l'attività collaborativa prevista nella seconda parte del percorso laboratoriale di Fisica.

Numerosi studi in letteratura ci forniscono un quadro teorico di riferimento per il nostro lavoro. Possiamo riassumere lo stato dell'arte osservando che i learning analytics inerenti il comportamento degli studenti in piattaforma hanno una grande utilità in ambito didattico. Ad esempio la frequenza dei login ad una piattaforma e-learning può essere utilizzata per la predizione del voto finale di uno studente (Froissard et al., 2015). In Jo ed altri (2014), viene dimostrato un modo efficace per utilizzare i dati relativi ai log degli utenti, come il tempo totale online e la frequenza dei login, come indicatori predittivi delle prestazioni di apprendimento. In Macfadyen (2010), viene dimostrato inoltre come l'uso di dati di tracciamento dello studente come, ad esempio, tempo speso online, login effettuati, file visti, web link e caricamento di esercizi in piattaforma (consegne sottomesse), sia correlato con il voto finale dello studente.

In Fig. 1 è rappresentato il flusso delle attività on-line, della ricerca, dello sviluppo del software e della creazione dei gruppi eterogenei.

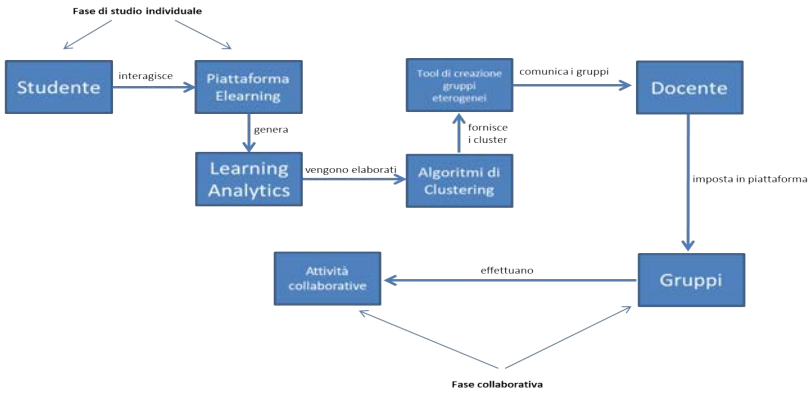

Fig. 1: Diagramma a blocchi del flusso delle attività, ricerca e sviluppo del software

Il processo per la formazione dei gruppi eterogenei ha richiesto due fasi di lavoro distinte:

a) implementazione di tecniche di clustering per la formazione di raggruppamenti omogenei in termini di caratteristiche e comportamenti simili degli studenti durante la fruizione del percorso individuale on-line;
b) sviluppo di un software automatico in grado di selezionare gli studenti dai diversi raggruppamenti omogenei per la formazione di gruppi eterogenei.

3.1. *Tecniche di clustering per la formazione di raggruppamenti omogenei*

Per la formazione di raggruppamenti omogenei abbiamo:

1. estratto e selezionato i learning analytics dalla piattaforma Moodle da utilizzare per la preparazione di un dataset da sottoporre alle tecniche di clustering;
2. implementato le tecniche di clustering ai LA per la realizzazione di raggruppamenti omogenei di studenti con caratteristiche e comportamenti simili tra loro.

Durante il primo step del progetto di ricerca abbiamo estratto, selezionandoli tra quelli disponibili, tutti i Learning Analytics relativi al comportamento degli studenti prodotti in piattaforma dopo la fruizione del percorso individuale online: frequenza login, ultimo login effettuato, tempo totale speso on-line, numero di video tutorial visti, frequenza video tutorial visti, numero video esperimenti visti, frequenza video esperimenti visti, numero pagine web viste, numero file pdf scaricati, numero di esercizi svolti. Per semplificare il file dei dati li abbiamo raggruppati in features, ovvero dati appartenenti allo stesso tipo di attività (per es. visualizzazione dei 5 video esperimenti) sono stati aggregati in una stessa feature. La feature, in italiano caratteristica, è una proprietà individuale e misurabile di un fenomeno osservato (Bishop, 2006).

Per la caratterizzazione degli studenti e la creazione dei raggruppamenti omogenei abbiamo utilizzato solo dati quantitativi, non prendendo in considerazione la valutazione degli elaborati da parte del docente per evitare che il peso del voto potesse influenzare il processo di suddivisione degli studenti. Le valutazioni degli elaborati individuali sono invece state utilizzate successivamente per verificare l'efficacia delle

tecniche di clustering. Abbiamo infatti messo a confronto le valutazioni con i cluster (raggruppamenti) ottenuti per vedere se esistesse una correlazione tra di essi, in modo da verificare se il comportamento degli studenti influisca o meno sulle loro performance finali.

Prima di applicare l'algoritmo per il clustering, i dati raccolti sono stati pre-processati. Ogni studente è stato rappresentato da un "vettore di input" con features costituite dai valori degli attributi associati allo studente.

Tutti i dati, organizzati in vettori di features, uno per ogni studente, sono stati inseriti all'interno di un unico file Excel, chiamato dataset, che rappresenta il file di input del nostro software per la generazione dei cluster. Per avere la stessa scala di valori i dati sono stati normalizzati in un intervallo che va da 0 a 1.

Successivamente siamo passati alla realizzazione del software per la creazione dei raggruppamenti omogenei utilizzando il linguaggio di programmazione Python poiché presenta una grande quantità di tool (librerie) per l'elaborazione di dati usando algoritmi di Machine Learning. Inizialmente il software prende in input il dataset e lo processa per determinare il numero di raggruppamenti da creare. Questo numero deve essere fornito all'algoritmo di clustering. Per generare i raggruppamenti è stato utilizzato l'algoritmo di clustering K-means, implementato attraverso una funzione della libreria scikit learn di Python (Hackeling, 2014). L'algoritmo utilizza un numero K di cluster (da impostare prima dell'esecuzione), per il quale si vuole suddividere il dataset e ad ogni cluster assegna un centroide, che deve rappresentare il punto centrale di ogni cluster.

Per determinare il numero adeguato di cluster è stata sviluppata una funzione in Python, utilizzando l'"Elbow Method" (Thorndike, 1953; Hackeling, 2014), un metodo di interpretazione all'interno dell'analisi del cluster finalizzato a trovare il numero appropriato di cluster in un set di dati. Il metodo permette di generare un grafico, avente nell'asse delle ascisse un intervallo di valori di K (ovvero il numero di cluster in cui viene suddiviso il dataset) e sull'asse delle ordinate la somma delle distanze dei dati osservati dai centroidi dei cluster, chiamato Within-Cluster-Sum-of-Squares (WCSS). Il numero di K (valore presente nell'asse delle ascisse) dove, nel grafico, la diminuzione del valore di WCSS all'aumentare di K comporta un calo significativo della velocità di incremento, viene chiamato "elbow". Tale valore rappresenta il numero ottimale di cluster da realizzare in base al dataset fornito.

Il passo successivo è l'esecuzione dell'algoritmo di clustering. L'algoritmo prende il numero K di cluster ottenuto dall'elbow method per suddividere il dataset e assegna ad ogni cluster un centroide. La scelta

del centroide, nella prima iterazione, avviene in maniera casuale. La vicinanza di un vettore (set di dati relative a un certo studente) al centroide di un cluster stabilisce l'appartenenza di tale vettore a quel determinato cluster. L'algoritmo calcola la distanza euclidea tra ogni vettore x e ogni centroide, assegnando il vettore al centroide c per il quale la distanza risulti minima:

$$c = \underset{c_i \in C}{\arg\min}\ dist(c_i,\ x)^2$$

dove "c_i" rappresenta un centroide dell'insieme C (insieme di centroidi), x rappresentano i vettori di input, mentre "dist" è la distanza euclidea standard. In seguito, vengono ricalcolati i valori dei centroidi. Il nuovo valore di un centroide sarà la media di tutti i vettori che sono stati assegnati al cluster del centroide. L'esecuzione continua per iterazione (un numero massimo di iterazioni vengono definite a priori per evitare un ciclo infinito) e termina quando:

- nessun vettore del dataset è soggetto a cambiamento di cluster;
- la somma delle distanze viene ridotta al minimo;
- viene raggiunto un numero massimo di iterazioni.

Alla fine dell'esecuzione vengono generati i raggruppamenti di studenti che risultano avere dei comportamenti e caratteristiche simili in piattaforma. In output viene generato il grafico con la rappresentazione dei vettori associati a diversi cluster in un piano bidimensionale di due tra le 10 features utilizzate ed il testo contenente il nome del raggruppamento, seguito da una lista di numeri, in cui ogni numero identifica ciascun studente.

Infine sono stati utilizzati i voti dell'elaborato individuale per poter vedere se risultano esserci delle somiglianze tra studenti, appartenenti ad uno stesso raggruppamento, non solo a livello di comportamento ma anche a livello di performance, per garantire una maggiore eterogeneità nella formazione dei gruppi eterogenei.

L'esecuzione del software continua con la formazione dei gruppi eterogenei, attraverso un algoritmo che preleva studenti da cluster diversi e li riorganizza in gruppi eterogenei.

3.2. *Formazione gruppi eterogenei*

Dopo la realizzazione dei cluster, l'elaborazione continua con l'esecuzione di una funzione Python basata su un nuovo algoritmo, sviluppato

appositamente, per suddividere in maniera automatica e uniforme gli studenti appartenenti a diversi cluster in diversi gruppi eterogenei.

Ogni raggruppamento di studenti ottenuto dal cluster rappresenta i diversi livelli di partecipazione alle attività in piattaforma e identifica tipologie di comportamento diverse per ogni cluster.

L'algoritmo prevede la selezione e l'inserimento all'interno di ogni gruppo di almeno uno studente appartenente a una tipologia diversa di cluster, in modo da garantire che in ogni gruppo siano presenti studenti con diverse caratteristiche e comportamento, in modo da migliorare così le performance degli studenti nelle attività collaborative.

Inizialmente la funzione calcola il numero di gruppi da creare in base al numero di studenti da inserire nel gruppo, che deve essere impostata a priori. Poi ordina le liste dei cluster in ordine crescente, dalla lista meno numerosa alla più numerosa e calcola, in base alle lunghezze, quanti studenti per ogni cluster inserire in un gruppo.

Successivamente l'algoritmo seleziona gli studenti in base al proprio ID, prelevando uno o più valori alla volta da ogni cluster (in base ai calcoli effettuati considerando le lunghezze dei cluster) in maniera sequenziale e li salva in un'unica nuova lista nell'ordine in cui essi sono stati prelevati dai diversi cluster. La lista ottenuta viene poi suddivisa in sotto liste attraverso la definizione di un intervallo che rappresenta il numero di partecipanti che si vuole ottenere all'interno di ogni gruppo.

In questo modo avremo la certezza che almeno un componente di ogni cluster (a patto che il numero di componenti dei gruppi non sia inferiore al numero di cluster) farà parte di ogni sotto lista (che rappresenta ogni singolo gruppo), ottenendo cosi al suo interno eterogeneità tra gli studenti.

Conclusa questa fase, una e-mail viene inviata direttamente al docente con la lista dei gruppi eterogenei contenenti al loro interno gli ID associati agli studenti.

4. Risultati e discussioni

Per la fase di test, è stato necessario creare il dataset, un file Excel da fornire in input al software, che presenta al suo interno diversi dati relativi al comportamento degli studenti. Sulla base del dataset fornito, il software, applicando l'algoritmo dell'"elbow method", ha restituito in output il relativo grafico dal quale è stato ottenuto il numero ideale di Cluster da generare, pari a 3 (Fig. 2).

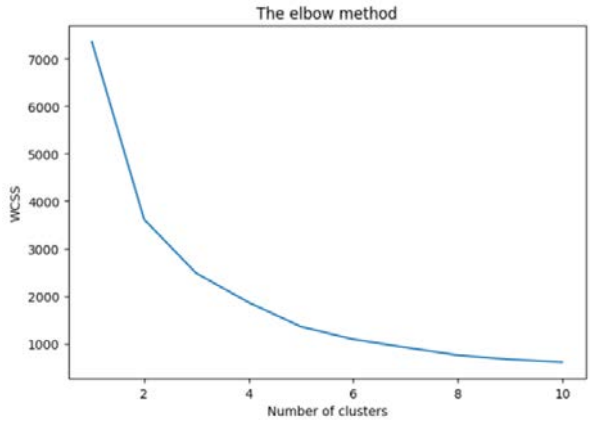

Fig. 2: Grafico elbow method

Successivamente, è stato eseguito l'algoritmo di clustering K-means che ha generato 3 raggruppamenti omogenei, caratterizzati rispettivamente da 10 (cluster 0 in rosso), 29 (cluster 1 in blu) e 16 (cluster 2 in verde) studenti (Fig. 3)

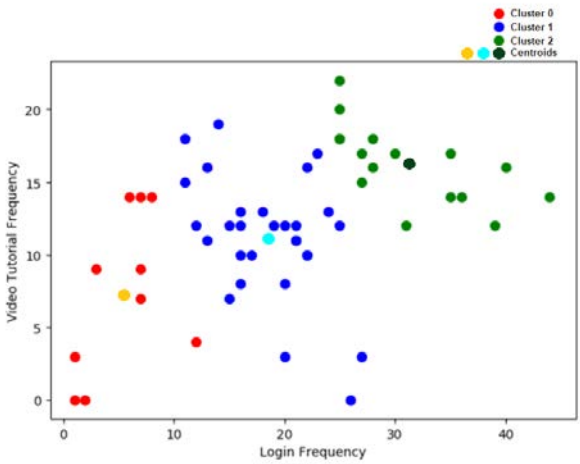

Fig. 3: Rappresentazione di una proiezione bidimensionale (frequenze di visualizzazione dei video tutorials, frequenza di login) dei raggruppamenti emersi dalla fase di clustering in base alle diverse features selezionate

I cluster ottenuti sono stati poi analizzati per determinare i diversi comportamenti in piattaforma, evidenziando quali features risultano essere più influenti all'interno dei raggruppamenti e in tal modo profilare gli studenti che appartengono a un determinato raggruppamento. I risultati sono i seguenti. Come si evince dalla Tab. 1, il cluster 0 rap-

presenta gli studenti poco attivi in piattaforma con un basso numero di accessi (media = 5,4), bassa frequenza di visualizzazioni relativi ai video tutorial (media = 7,4), e uno scarso numero e frequenza di visualizzazioni relativa agli esperimenti (media = 2,6). La scarsa partecipazione si riflette anche sulla consegna dell'elaborato individuale, dove solo il 40% degli studenti lo ha consegnato in piattaforma. Il cluster 1 rappresenta gli studenti mediamente attivi in piattaforma, in cui gli utenti hanno ottenuto dei valori medi in quasi tutte le attività: media del numero di accessi pari a 18,4, media del numero di video tutorial visti uguale a 2,5 media della frequenza di visualizzazioni relativi ai video tutorial uguale a 11,2, media del numero di esperimenti visti uguale a 4,1, media della frequenza di visualizzazione dei video esperimenti uguale a 6,7. In questo caso la partecipazione ha consentito al 90% degli studenti di realizzare e consegnare l'esercizio finale. Il cluster 2 invece rappresenta gli studenti molto attivi in piattaforma con una media del numero di accessi di 31,2, media del numero di video tutorial visti uguale a 3, con una frequenza media di visualizzazioni pari a 16,2. La media del numero di esperimenti visti è di 4,3 (5 video esperimenti in totale) mentre la frequenza media di visualizzazioni è di 7. Questo livello di partecipazione ha consentito a tutti gli studenti di consegnare l'elaborato finale (100% degli studenti).

	cluster 0	**cluster 1**	**cluster 2**
	Studenti poco attivi in piattaforma	Studenti mediamente attivi in piattaforma	Studenti molto attivi in piattaforma
n° studenti	10	29	16
n° accessi in piattaforma (media)	5,4	18,4	31,2
n° video tutorial visti (media)	1,8	2,5	3
frequenza visualizzazione totale dei video tutorial (media)	7,4	11,2	16,2
n° video esperimenti visti (media)	2	4,1	4,3
frequenza visualizzazione totale dei video esperimenti (media)	2,6	6,7	7
% studenti che hanno svolto l'elaborato	40%	90%	100%

Tab. 1: Differenze tra cluster e dettagli delle features analizzate

Per verificare se il comportamento svolto dagli studenti nella piattaforma e-learning potesse avere un riscontro in termini di performance, abbiamo confrontato ogni cluster con la media dei voti relativi all'elaborato individuale svolto dagli studenti appartenenti al medesimo cluster. I risultati ottenuti sono molto interessanti. Come riportato in Fig. 4, il livello di partecipazione in piattaforma degli studenti rispecchia pienamente il voto finale (voto in centesimi). Gli studenti più attivi (cluster 2), infatti, hanno ricevuto un punteggio più alto rispetto agli altri con una media dei voti pari ad 82,5. Gli studenti mediamente attivi (cluster 1) risultano invece avere una votazione inferiore con una media del 72,04, mentre quelli del cluster 0 una media dei voti addirittura al di sotto della sufficienza, pari a 53. Questa forte correlazione evidenziata dall'indice di Pearson pari a 0,986 ci ha permesso di confermare ulteriormente le differenze esistenti tra gli studenti appartenenti ai diversi cluster, non solo più in termini di comportamento ma anche di performance.

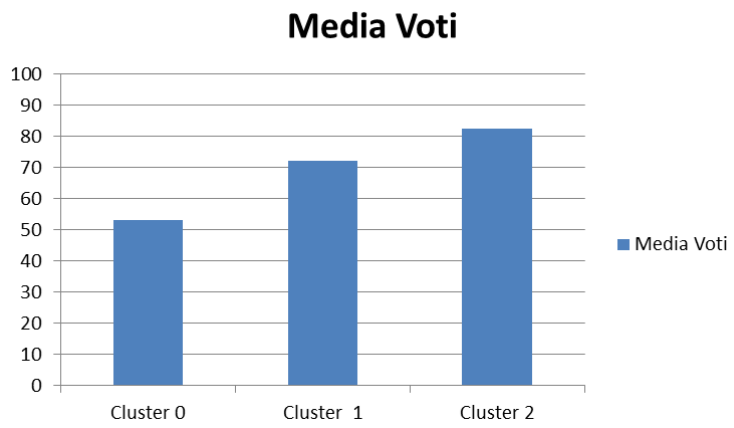

Fig. 4: Medie dei voti (intervallo 0-100) degli studenti nei diversi cluster

Una volta ottenuti i cluster, l'esecuzione del software ha restituito la creazione dei gruppi eterogenei. A priori abbiamo impostato il numero di studenti per gruppo pari a 5. Dalla Fig. 5 possiamo osservare che una volta impostato l'algoritmo, il software ha creato 11 gruppi e distribuito in automatico gli studenti in modo che in ogni gruppo fosse presente almeno uno studente appartenente a cluster differenti, creando quindi gruppi eterogenei sulla base dei comportamenti degli studenti.

```
cluster 0:
[4, 8, 12, 22, 25, 29, 30, 34, 43, 46]
cluster 1:
[0, 1, 2, 3, 5, 6, 7, 10, 11, 14, 15, 21, 23, 24, 26, 27, 28, 31, 33, 35, 36, 38, 39, 40, 41, 45, 47, 51, 52]
cluster 2:
[9, 13, 16, 17, 18, 19, 20, 32, 37, 42, 44, 48, 49, 50, 53, 54]


Gruppo 0[4, 9, 0, 1, 2]
Gruppo 1[8, 13, 3, 5, 6]
Gruppo 2[12, 16, 7, 10, 11]
Gruppo 3[22, 17, 14, 15, 21]
Gruppo 4[25, 18, 23, 24, 26]
Gruppo 5[29, 19, 27, 28, 31]
Gruppo 6[30, 20, 33, 35, 36]
Gruppo 7[34, 32, 38, 39, 40]
Gruppo 8[43, 37, 41, 45, 47]
Gruppo 9[46, 42, 44, 51, 52]
Gruppo 10[44, 48, 49, 50, 53]
```

Fig. 5. Cluster e gruppi eterogenei ottenuti dall'esecuzione del software

Infine il software ha inviato automaticamente una e-mail al docente, con la lista dei gruppi eterogenei, indicando il nome del gruppo (es. gruppo 0, gruppo 1, etc.) e accanto ad ogni gruppo la lista di 5 numeri, che sono gli ID che identificano gli studenti. In tal modo il docente ha potuto selezionare gli studenti in base al proprio ID, inserirli nei relativi gruppi all'interno di Moodle e iniziare la parte collaborativa del corso.

4. Conclusioni e sviluppi futuri

In questo lavoro abbiamo ideato e realizzato un nuovo software finalizzato ad aiutare i docenti nella definizione di gruppi di studenti destinati ad attività collaborative. La possibilità di utilizzare tecniche di Machine Learning ci ha permesso di dimostrare come queste contribuiscano al miglioramento della formazione di gruppi eterogenei, ma anche di testare l'efficienza dell'algoritmo K-means, analizzando per ogni cluster le features e delineando diversi profili di studente. Il confronto tra i cluster ottenuti e la media del voto finale degli studenti di ogni cluster, ha confermato le differenze tra i cluster anche in termini di performance. Grazie all'esecuzione di queste tecniche sui dati raccolti e le verifiche sull'efficienza delle tecniche, è stato possibile garantire e massimizzare l'eterogeneità degli studenti all'interno di ogni gruppo, accrescendo così le possibilità di successo per tutti gli studenti nello svolgimento di lavori di gruppo all'interno del corso. La scelta delle features è ricaduta sui Learning Analytics della piattaforma e-learning Moodle, per permettere al software di essere riutilizzato da altri docenti e tutor che svolgono attività didattiche on-line su Moodle. La scelta dell'ambiente Moodle non risulta comunque essere vincolante per il funzionamento del software, in quanto è facilmente adattabile a qualsiasi altro LMS (Learning Management System), attraverso la modifica

del dataset. Il software realizzato per il presente lavoro richiede attualmente, per l'esecuzione, un supporto tecnico ai docenti, soprattutto nella realizzazione del dataset e per la scelta del numero di cluster da creare. Il progetto di ricerca proseguirà con la realizzazione di un Plugin avanzato per Moodle che permetterà al docente in modo semplice, intuitivo e automatizzato di creare gruppi di studenti eterogenei, attraverso la sola selezione, tramite delle caselle di spunta, dei Learning Analytics di Moodle che il docente stesso riterrà maggiormente rilevanti per la profilazione delle caratteristiche degli studenti.

Riferimenti bibliografici

Abedin B., Daneshgar F., D'Ambra J. (2012). Pattern of non-task interactions in asynchronous computer-supported collaborative learning courses. *Interactive Learning Environments*, 22 (1), 1-17.

Amendola D., Miceli C. (2018). Online peer assessment to improve students' learning outcomes and soft skills. *Italian Journal of Educational Technology*, 26(3), 71-84.

Bacon R.D., Stewart K.A., Anderson E.S. (2002). Methods of Assigning Players to Teams: A Review and Novel Approach. *Simulation & Gaming*, 32(1), 6-17.

Bishop C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Bovo A., Sanchez S., Héguy O., Duthen Y. (2013). "Clustering Moodle data as a tool for profiling students". The 2nd International Conference on e-Learning and e-Technologies inEducation - ICEEE (pp. 121-126).

Burke A. (2011). Group Work. *Journal of Effective Teaching*, 11(2), 87-95.

Csernica J., Hanyka M., Hyde D., Shooter S., Toole M., Vigeant M. (2002). *Practical guide to teamwork, version 1.1*. Lewisburg: College of Engineering, Bucknell University.

Dutt A., Ismail M. A., Herawan T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.

Feng M., Heffernan N.T. (2005). *Informing teachers live about student learning: Reporting in the assistment system*. The 12th Annual Conference on Artificial Intelligence in Education Workshop on Usage Analysis in Learning Systems. Amsterdam: IOS Press.

Foote E. (2009). *Collaborative Learning in Community College*. ERIC. Estratto da http://www.ericdigests.org /1998-1/ colleges.htm.

Froissard C., Richards D., Atif A., Liu D.Y. (2015). *An enhanced learning analytics plugin for Moodle: student engagement and personalised intervention*. Proceedings of the ASCILITE Conference, (pp. 180-189).

Hackeling G. (2014). *Mastering Machine Learning with scikit learn*. Birmingham: Packt Publishing

Jackson S.E., May K.E., Whitney K. (1995). Understanding the dynamics of diversity in decision making teams. In Guzzo & Salas (Eds), pp. 204-261.

Jo I., Kim D., Yoon M. (2014). *Analyzing the log patterns of adult learners in LMS using learning analytics*. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge (pp.183-187).
Kotsiantis S.B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. Proceedings of the Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, (pp. 3-24).
Lopèz M., Luna J., Romero C., Ventura S. (2012). *Classification via clustering for predicting final mark based on student partecipation in forums*. Proceedings in the International Conference on Information Technology Systems and Innovation (ICITSI), (pp.148-151).
Macfadyen L.P., Dawson S. (2010). Mining LMS data to develop an ''early warning system" for educators: A proof of concept. *Computer and Education*, 54, 588-599.
Maina E.M., Oboko R.O., Waiganjo P.W. (2017). Using Machine Learning Techniques to Support Group Formation in an Online Collaborative Learning Environment. *International Journal of Intelligent Systems and Applications*, 3, 26-33.
McLoughlin C., J.W. Lee M. (2007). *Social software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era*. Proceedings in Ascilite, Australian Society for Computers in Learning in Tertiary Education Annual Conference (pp. 664-675).
Nijstad B.A., De Dreu C. K.W. (2002). Creativity and Group Innovation. *Applied Psychology: An International Review*, 51 (3), 400-406.
Papamitsiou Z., Economides A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature. Review of Empirical Evidence. *Educational Technology & Society*, 17, 49-64.
Razmerita L., Brun A. (2011). *Collaborative learning in heterogeneous classes. Towords a Group Formation Methodology*. Proceedings of 3rd International Conference on Computer Supported Education (CSEDU 2011), 2, (pp. 189-194).
Sadeghi H., Kardan A.A. (2016). Toward effective group formation in computer-supported collaborative learning. *Interactive Learning Environments*, 24(3), 382-395
Smith B.L., MacGregor J.T. (2009). *What is collaborative learning? National Center on Postsecondary Teaching*. Pennsylvania State University Press.
Stevens D.D., Levi A.J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning*. Stylus Publishing.
Thorndike R.L. (1953). Who Belongs in the Family? *Psychometrika,* 18, 267-276.
Vellido A., Castro F., Nebot A. (2010). Clustering educational data. *Handbook of educational data mining*, 75-92.