Analisi Predittiva sul Dataset OULAD

Seminario di Cultura Digitale

Lavinia Rotellini

22/07/2025

Abstract

Lo scopo di questa analisi è proporre una breve ricognizione dello stato dell'arte nell'ambito delle learning analytics, procedendo poi con uno studio di analisi predittiva sullo Open University Analytics Dataset. La task affrontata nell'analisi è una classificazione multiclasse su una delle variabili del dataset: `final_result`. Questa può assumere tre diversi valori: *pass*, se lo studente ha superato con successo l'esame, *fail*, se lo studente non ha superato l'esame, oppure *withdrawl*, se lo studente si è ritirato dal corso. Infine, si procede con una breve analisi di explainability, dove si esplorano le features che hanno portato i modelli ad effettuare le scelte di classificazione: questo con l'obbiettivo di valutare quali possano essere gli aspetti più decisivi nell'andamento della carriera di uno studente.

1 Introduzione

Questo progetto è ispirato al seminario tenuto da parte di Daniela Rotelli, ricercatrice attualmente presso la Sorbonne Université, tenuto durante l'anno accademico 2023-2024 per il corso *Seminario di Cultura Digitale*. La presentazione verteva sull'utilizzo delle learning analytics allo scopo di migliorare la qualità dell'istruzione. Il settore delle learning analytics appartiene all'ambito del **Technology-Enhanced Learning** (TEL), termine utilizzato per indicare l'applicazione di tecniche dell'informazione e della comunicazione all'insegnamento e all'istruzione [13]. Questo settore si articola in tre ambiti principali:

- L'Educational Data Mining: si occupa di estrarre informazioni rilevanti da grandi insiemi di dati;
- Le Learning Analytics: finalizzato a potenziare l'apprendimento online;
- L'Academic Analysis: si focalizza sul migliorare l'apprendimento ed i risultati scolastici a livello nazionale od internazionale. [7]

Il TEL si trova dunque all'intersezione di molte discipline differenti: la business intelligence, le web analytics, il data mining e machine learning, ed i recommendation systems. L'aspetto fondamentale che ha portato allo sviluppo del TEL è connesso alla diffusione dell'approccio data-driven nei settori dell'intelligenza artificiale e dell'analisi di dati, approccio che ha potuto diffondersi grazie alla sempre maggiore mole di dati reperibili dal web. I *big data* infatti, insiemi di dati di dimensioni talmente considerevoli da non poter essere analizzati con i classici strumenti della statistica e dell'informatica, ma che richiedono approcci automatizzati e specifici per la loro gestione e analisi, hanno permesso di svolgere studi a scale che prima non erano immaginabili. Il Technology Enhanced Learning nasce già nella seconda metà degli anni '90, ed i primi studi in ambito di Educational Data Mining e Learning Analytics cominciano ad affermarsi nella prima decade degli anni 2000. In particolare, la prima conferenza di Educational Data Mining si è svolta a Montreal (Canada) nel 2008, mentre la prima conferenza di Learning Analytics è stata inaugarata a Banff (Canada) nel 2011[20]. Da questo momento iniziale, entrambi i campi hanno guadagnato in popolarità, con la nascita di importanti riviste come il *Journal of Educational Data Mining* ed il *Journal of Learning Analytics*, e di volumi sulle tecniche specifiche applicabili a questo ambito (primo fra tutti, *Data Mining in Education*[19]) In figura 1, il numero di papers ed eventi principali nell'ambito EDM/LA registrati su Google Scholar fino al 2019[20]. I dati utilizzati per entrambe queste discipline sono reperiti grazie alla diffusione di nuove metodologie di

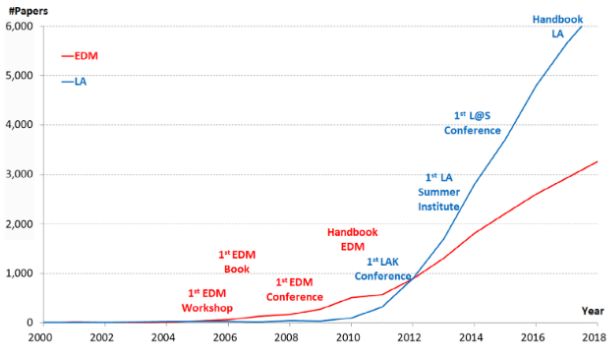

Figure 1: Numbers of Papers and Events in TEL

insegnamento, in particolar modo l'**apprendimento online** e i **Virtual Learning Environments** (VLE), che hanno cominciato a diffondersi alla fine del secolo scorso guadagnando in popolarità molto velocemente nei primi anni 2000.
Questo progetto si situa nell'ambito dell'educational data mining, "paradigma orientato al design di modelli, task, metodi e algoritmi per l'esplorazione dei dati di tipo educativo, così da scoprire schemi e realizzare predizioni che caratterizzino i comportamenti e i risultati degli studenti, dei contenuti di domain knowledge etc." [16].
Più nello specifico, il presente lavoro si colloca nell'ambito della knowledge discovery with models, una linea di ricerca che si avvale di modelli preesistenti per l'identificazione di relazioni tra i comportamenti degli studenti e i contesti educativi in cui essi operano [20]. A tal fine, saranno adottate tecniche di data mining e machine learning per la realizzazione di un'analisi predittiva dell'esito di un corso, basata su variabili demografiche e indicatori di engagement degli studenti.

2 Il Dataset

2.1 Aspetti problematici nella realizzazione di un EDM Dataset

Realizzare un dataset allo scopo di performare analisi di educational data mining è un compito piuttosto complesso a causa della forte granularità di situazioni e metodologie che possono essere applicate. Il primo passo approcciandosi ad un lavoro nel contesto dell'education data mining è comprendere appieno la task da svolgere. Bakhshinategh et al.[2] individuano una tassonomia composta da cinque famiglie di applicazioni nell'educational data mining:

- **Student Modeling**: processo che individua aspetti cognitivi delle attività realizzate dagli studenti, come l'analisi della performance e dei comportamenti degli studenti e l'individuazione di comportamenti indesiderabili;
- **Decision Support Systems**: lo sviluppo di applicazioni che permettano di migliorare il processo di istruzione semplificando le decisioni degli stakeholders;
- **Adaptive Systems**: utilizzo di sistemi intelligenti per adattare l'apprendimento all'utente;
- **Evaluation**: sviluppo di metodologie e sistemi che assistano gli insegnanti nella valutazione;
- **Scientific Inquiry**: investigare teorie sull'apprendimento e istruzione.

I dati possono essere poi estratti da diversi tipi di ambienti educativi: il più tradizionale è senza dubbio il face-to-face contact tra studenti ed insegnanti, che permette di analizzare aspetti come la presenza scolastica, i voti, gli obbiettivi del curriculum etc. La raccolta di dati in questo contesto è però molto difficile e prolungata nel tempo, dunque si preferisce seguire la strada della Computer-based Education e dei Web-based Educational Systems, piattaforme e sistemi che consentono l'utilizzo di computer e applicazioni web

per guidare, istruire e seguire lo studente[19].
I Web-based Educational Systems di maggiore utilizzo rientrano nelle seguenti categorie:

- **Learning Management Systems**: applicazioni software che automatizzano l'amministrazione, il tracking e il report di eventi all'interno di un ambiente di istruzione[6];
- **Massive Open Online Courses**: corsi online tenuti per un numero non limitato di partecipanti, da professori o altri esperti [29];
- **Intelligent Tutoring Systems**: ambienti di apprendimento computerizzati che incorporano modelli computazionali originari delle scienze cognitive, dell'apprendimento, dalla linguistica computazionale e dall'intelligenza artificiale. Riescono a tracciare lo stato psicologico dell'apprendente nel dettaglio attraverso un processo chiamato *student modeling*[8];
- **Adaptive and Intelligent Hypermedia Systems**: sistemi che costruiscono un modello dell'obbiettivo, delle preferenze e della conoscenza di ogni user, così da adattare l'ipertesto a sua disposizione alle specifiche necessità, ipertesto che può essere qualsiasi supporto utilizzato per l'apprendimento, da una presentazione ad un'enciclopedia[3];
- **Virtual Learning Environments**: sistema software creato per supportare l'insegnamento e l'apprendimento permettondo ad insegnanti e studenti di interagire in un sistema online integrato;
- **Quiz Systems**: ovvero quiz effettuati con l'utilizzo di apparecchiature elettroniche.

I dati estraibili utilizzando queste risorse possono appartenere a categorie diverse. La più diffusa è certamente la categoria dei *Relational Data*, insiemi di databases con un identificatore univoco, che contengono tuple di oggetti identificabili tramite una chiave e descritti da una serie di attributi. Possiamo individuare poi i *Transactional Data*, composti da una singola tabella dove ogni record rappresenta una transazione identificata con un numero ed una lista di elementi che ne fanno parte. Successivamente si distingue tra *Temporal, Sequence and Time Series Data*, ovvero sequenze di valori o eventi che cambiano nel tempo, *Text Data*, ampie collezioni di documenti da fonti diverse, come articoli di notizie, libri e pagine web. Fondamentali sono poi i *Multimedia Data*, i cui database contengono dati in formato di immagine, audio e video, molto utilizzati attualmente per la realizzazione dei large language models multimodali, ed infine i *World Wide Web Data*, termine usato in riferimento ai dati estraibili dai contenuti e dalla struttura delle pagine web oltre che alle interazioni degli utenti.

2.2 Specificità del Preprocessing

Il data mining applicato all'istruzione segue tendenzialmente il processo di Knowledge Discovery in Databases (KDD), secondo il quale il preprocessing dei dati deve seguire fasi predeterminate. La prima di esse è la fase di raccolta dei dati, nella quale si definisce l'obbiettivo dello studio e si procede con la ricerca dei dati più adatti a perseguirlo. La seconda fase riguarda la data understanding: per sfruttare al meglio i dati raccolti è necessario conoscerli, dunque si procede con l'esplorazione delle loro origini, della composizione, se ne osservano le distribuzioni statistiche e come essi interagiscono tra loro. Si passa poi al sampling, che consiste nella selezione degli attributi di maggior interesse per la task, e poi la trasformazione, identificazione e rimozione del rumore: Il rumore, assumendo che i dati regolari provengano da una certa distribuzione, comprende tutte le istanze che non sono generate da tale distribuzione. La fase successiva è la normalizzazione, che riporta i dati sulla stessa scala, ed infine la feature engineering, fase nella quale si passa alla creazione, selezione e trasformazione delle variabili (o "feature") utilizzate per costruire modelli predittivi o descrittivi a partire dai dati grezzi.
In ambito di educational data mining esistono, oltre a queste fasi, tecniche e accorgimenti specifici che permettono di estrarre informazioni dai learning environments in maniera accurata. La questione è stata affrontata in *A Survey on Pre-Processing Educational Data*[19], dove sono identificate otto fasi di preprocessing e proposti una serie di tools per svolgerle.

Il processo di preprocessing inizia con la **data gathering**, o fase di raccolta dei dati. La maggior parte dei tools e sistemi di apprendimento di solito cattura le azioni che gli studenti fanno con le dita, questo

significa che è registrato ogni click del mouse o della tastiera, i quali sono poi restituiti in file di registro come lista ordinata di eventi. I dati contenuti nel registro, oltre alle azioni degli studenti, contengono spesso anche data ed ora in cui si è verificata l'interazione, l'indirizzo IP da cui è partita la richiesta al server del VLE, il metodo ed il nome della richiesta. Alcuni di essi registrano anche lo user name e informazioni aggiuntive sull'azione compiuta, come la durata.
Dati relativi a quiz e test sono invece organizzati in matrici che riportano l'identificatore dello studente, gli oggetti su cui è stato testato ed il relativo punteggio.
Infine, possiamo individuare i portfoli, registri di apprendimento che restituiscono il profilo dello studente con le sue attività, le discussioni, i report dell'apprendimento online, il tipo di apprendimento preferito, l'anno di frequenza e la difficoltà dei corsi.

Si passa poi all'**aggregazione ed all'integrazione dei dati**. Poiché i dati sono raccolti da molti contesti diversi, è necessario riunirli in maniera coerente in un unico database, così da facilitare notevolmente il lavoro. È inoltre possibile che i dati non siano tutti disponibili digitalmente, basti pensare alle performance scritte dello studente o al registro delle presenze: questa fase può dunque implicare anche un lungo processo di digitalizzazione delle risorse.
La scelta della struttura per la gestione dei dati dipende poi dalla loro mole e dalle esigenze specifiche, i *relational databases* sono sicuramente molto potenti e flessibili, ma si può optare anche per *tabelle Pivot* in Excel, o per gli *archivi informatici*, o infine per i *data cube structures*, combinazioni di array di dati che permettono di visualizzare diversi attributi contemporaneamente.

La task successiva è la **pulizia dei dati**, nella quale dati erronei o non rilevanti vengono eliminati. Il primo tipo di dato problematico da affrontare sono i valori mancanti, la cui strategia di riempimento dipende dalle esigenze dello studio: potremmo sostituire il valore mancante con la media o la mediana degli altri valori della colonna, oppure utilizzare un metodo di regressione per approssimarne il valore. In un contesto educativo, valori di questo tipo appaiono in genere quando gli studenti non hanno completato tutte le attività in un corso: se il numero di valori mancanti è molto elevato possiamo anche decidere di eliminare l'istanza.
Successivamente, è necessario stabilire la strategia per gestire i valori outlier. Questi possono essere dovuti ad errori di misurazione oppure all'inerente variabilità dei dati. Secondo la procedura del knowledge discovery in databases, gli outlier dovrebbero essere eliminati, ma nell'ambito dell'istruzione performance anomale da parte di studenti o studenti con caratteristiche particolari possono essere casi reali, è dunque fondamentale conoscere il dominio in cui i dati sono stati estratti per comprendere come gestire queste osservazioni.
Infine, è possibile che siano registrate inconsistenze nei dati per le quali certi gruppi sono diversi da altri senza un'apparente motivazione: queste situazioni possono essere imputate spesso a registrazioni erronee.

Si passa poi alla **User and Session Identification**. Questa fase è particolarmente delicata, poiché qualsiasi informazione identificativa dello studente è protetta da regolamenti sulla privacy, il che implica che è necessario anonimizzare aspetti come lo user ID o informazioni personali utilizzando numeri generati incrementalmente o altre eventuali strategie richieste.

Si procede poi con la **variable selection** e la **data filtering**, che consistono nella scelta degli attributi più rilevanti all'obbiettivo prefissato: è necessario che i dati siano compresi e preprocessati così da distinguere al meglio quali variabili contengono più informazione utile rispetto ad altre che possono essere superflue o ridondanti. In ambito di istruzione, i dati disponibili variano molto per granularità ed interesse: dai VLE è possibile estrarre informazioni utili come i voti registrati da uno studente o il luogo di provenienza, ma anche aspetti meno fondamentali come la sua mail o numero di telefono, che non ci forniscono informazioni riguardo il suo percorso accademico e che possono essere scartate.

I dati rimasti dopo le fasi precedenti possono poi essere modificati durante la fase di **data transformation** usando tecniche di normalizzazione, discretizzazione, derivazione e conversione di formato. La discretizzazione è un aspetto particolarmente importante in questo ambito, poiché variabili molto specifiche e dettagliate possono essere condensate in nuove features maggiormente informative.

Tenere conto delle peculiarità del dato in ambito educativo è dunque fondamentale per un'analisi efficace.

2.3 I Dataset per l'Educational Data Mining e le Learning Analytics

L'analisi condotta in questo lavoro ha richiesto il supporto di dati empiricamente validi, accuratamente documentati e fondati su basi metodologiche solide, al fine di garantire un approccio rigorosamente data-driven. A tal fine, è stata effettuata una ricognizione sistematica dello stato dell'arte relativa ai dataset disponibili, che ha consentito di individuare un insieme di opzioni interessanti. In particolare, si segnala la ricognizione effettuata nell'articolo *Review on publicly available datasets for educational data mining*[17].
Il paper individua cinque fonti principali di dati in ambito di istruzione, che procediamo a presentare.

La repository maggiormente sedimentata dove è possibile reperire dati in ambito EDM è la **UCI ML Repository** [22], che contiene alcuni tra i dataset più citati in letteratura. Alcuni esempi sono l'*Educational Process Mining Dataset*[25], che contiene dati di attività effettuate da studenti che interagiscono con un educational simulator, l'*Open university Learning Analytics Dataset*[14], dataset sedimentato in ricerca la cui problematica fondamentale sono le ampie dimensioni e la struttura in sette file csv che richiedono intenso preprocessing, e lo *Student Academics Performance Data Set*, che contiene per lo più informazioni di tipo demografico.
Un'altra repository di rilievo è la **Mendeley Data Repository**[23], che presenta dataset piuttosto nuovi e ancora non di uso sedimentato, come il *Dataset for Factors Assessing Teacher's Burnout*, con dati estratti grazie all'utilizzo di questionari, e l'*Embeddings and Topic Vectors for MOOC Lectures Dataset*, che presenta word embeddings e distribuzioni di document topic generati dalle trascrizioni di 12,032 corsi provenienti dai massive online open courses.
Si segnala inoltre l'**Harvard Dataverse**[24], organizzato per argomenti, che mette a disposizione dataset interessanti come l'*HarvardX Person-Course Academic Year 2013 De-identified dataset*[18], raccolta di dati statistici dettagliati di corsi online e il *Early Reading and Writing Assessment in Preschool Using Video Game Learning Analytics*[1], che presenta osservazioni e valutazioni sulle capacità precoci di lettura e scrittura attraverso i videogiochi.
Si distingue inoltre la **DataSchop@CMU data source**[4], sistema che raccoglie dataset dettagliati e longitudinali che permettono di fare considerazioni nel tempo.
Infine, possiamo segnalare i dataset realizzati per competizioni in ambito EDM o provenienti da piattaforme come Kaggle. In particolare, quelli risultanti dalla *KDD Cup 2010 Educational Data Mining Challenge*[11] e dalla *KDD Cup 2015*[12], oppure il *Kaggle: Students Performance in Exams*[9] ed il *Duolingo Shared Task on Second Language Acquisition Modeling*[5].

2.4 Il Dataset OULAD

Il dataset scelto per l'analisi è l'**Open University Learning Analytics Dataset**[14] o **OULAD**. La scelta è virata su questo dataset per tre ragioni fondamentali. Innanzitutto, durante la fase di ricerca dei dati, l'OULAD è emerso come uno tra i contendenti meglio realizzati, completi, e maggiormente conosciuti, il cui uso è già sedimentato in letteratura. La seconda ragione che lo ha spinto ad essere scelto come riferimento per questo studio è la sua accessibilità, sia in termini di reperibilità che di utilizzo. Infine, presenta sette tabelle riferite ad ambiti diversi del processo educativo, dunque una varietà di informazioni che permette un ampio ventaglio di analisi, sia per studente che per corso.

Il dataset raccoglie dati sulla struttura di 40 corsi e sugli studenti e le loro interazioni con il VLE della Open University (OU) di Londra, risalenti agli anni 2013 e 2014. La OU è una delle più grandi università a distanza nel mondo, al 2023 contava un totale di 199,400 iscritti[26].
I corsi sono chiamati `moduli` e possono essere presentati più volte in un anno accademico. All'interno del dataset, al fine di segnalare il periodo temporale in cui si è tenuto il corso, è utilizzato l'anno seguito da una lettera che indica il mese di appartenenza: dalla A per gennaio fino alla J di ottobre.
Il dataset è student-oriented, il che significa che il punto focale è lo studente, le cui informazioni personali sono state completamente anonimizzate.

I dati sono stati estratti dalla OU data warehouse, che aggrega informazioni da ognuno dei diversi sis-

temi di raccolta informazioni sugli studenti e sui corsi, e ne è stato selezionato un campione rappresentativo seguendo alcue regole:

- Il numero di studenti iscritti al corso è maggiore di 500
- Esistono almeno due presentazioni del modulo
- I dati del corso sono reperibili tramite il VLE
- Il modulo ha un numero significativo di studenti bocciati

Si distinguono tre diversi tipi di dato: le informazioni demografiche, che rappresentano le informazioni d base riguardo gli studenti, come l'età, il genere e la regione di provenienza. Abbiamo poi i dati sulla performance, che riflettono i risultati degli studenti, ed infine i dati sul metodo di apprendimento, che fanno riferimento a tutte le attività dello studente sul VLE.

3 L'analisi

3.1 Merge e Semantica dei Dati

Il dataset OULAD raccoglie i dati in sette file in formato `comma separated values`. All'interno di questi file, i dati sono presentati in tabelle chiamate 'databases', poiché ne presentano la struttura. Queste tabelle sono:

- **studenInfo**: contiene le informazioni demografiche riguardo gli studenti ed i risultati che hanno ottenuto nel modulo studiato. In totale presenta 32'593 righe;
- **courses**: contiene la lista di tutti i moduli disponibili con le rispettive presentazioni. Presenta in totale 32 righe;
- **studentRegistration**: contiene informazioni riguardo quando lo studente si è registrato per la presentazione del modulo, registrando in totale 32'593 righe;
- **assessment**: contiene informazioni riguardo tutti gli assessment per ogni modulo. Ognuno di essi contiene infatti ogni assessment (traducibile come test intermedio) seguito da un esame, presenta 206 righe;
- **studentAssessment**: contiene i risultati per ogni prova intermedia per ogni studente eccezion fatta per l'esame, il cui voto è incluso nello score finale. Presenta 173'912 righe;
- **studentVle**: contiene informazioni riguardo alle interazioni dello studente con il virtual learning environment. Poichè registra ogni singola interazione, presenta 10'655'280 righe;
- **vle**: contiene informazioni riguardo alle informazioni relative ai materiali disponibili sul VLE.

Fin da subito è emersa la necessità di lavorare con un unico dataset, il che implicava unire le tabelle in un unico dataframe. Lo strumento utilizzato per fare ciò è stato `SQLite`, che con il comando `merge` permette di aggregare due tabelle in una, optando per una tra due diverse metodologie: **inner join**, mantiene solamente le righe perfettamente identiche tra le due tabelle sulla base della chiave specificata, oppure **outer join**, che mantiene tutte le righe. La scelta è ricaduta sull'inner join, poiché l'informazione principale di nostro interesse era l'id_student, dato che la classificazione è da applicare ad ogni studente e ad esso soltanto e questo implicava mantenere ognuna delle diverse chiavi identificative. La tabella di più difficile gestione si è presentata fin da subito essere `studentVle`, a causa delle quasi 11 milioni di istanze che conteneva. Ad una prima esplorazione è risultato subito chiaro però che alcune informazioni al suo interno fossero di importanza fondamentale: in particolar modo, l'attributo `sum_click`, che rappresentava il numero di interazioni dello studente con il materiale a sua disposizione, e si presentava come un forte indicatore di engagement molto utile per il task di classificazione. Si è dunque proceduto a creare una nuova variabile chiamata `total_clicks`, risultante dalla somma di tutti i click dello studente relativi ad una determinata presentazione del corso.

Questo ha permesso di ridurre notevolmente il numero di entrate e di aggregare `studentVle` con `studentData` sulla base di due diverse features: `code_presentation` e `id_student`, risultando in una tabella chiamata `student_data` di 29369 righe e le seguenti features, sia di tipo qualitativo (features categoriche) che quantitativo (features numeriche):

code_module Codice del modulo di studio (es. AAA).

code_presentation Codice della presentazione del modulo (es. 2014J).

id_student Identificativo univoco dello studente.

date_registration Giorno (numero relativo) di registrazione al modulo.

date_unregistration Giorno in cui lo studente si è disiscritto (se presente).

gender Genere dello studente (M/F).

region Regione geografica di provenienza dello studente.

highest_education Livello di istruzione più alto raggiunto.

imd_band Indice di deprivazione socio-economica (es. 0-10

age_band Fascia d'età dello studente (es. 35-55).

num_of_prev_attempts Numero di tentativi precedenti dello stesso modulo.

studied_credits Numero di crediti già studiati prima di questo modulo.

disability Indicatore di disabilità (Y/N).

final_result Risultato finale nel modulo (Pass, Fail, Withdraw, Distinction).

module_presentation_length Durata della presentazione del modulo (in giorni).

total_clicks Numero totale di clic effettuati sulla piattaforma VLE.

Si è proceduto successivamente con il merge delle restanti tabelle, che hanno portato la versione finale di `student_data` ad un totale di 173'744 entrate per 23 colonne. Le features aggiunte in questa fase sono le seguenti:

id_assessment Identificativo univoco dell'assessment (compito, esame, quiz, ecc.).

assessment_type Tipo di valutazione, se ne distinguono tre tipi: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).

date Giorno relativo alla presentazione dell'assessment nel modulo.

weight Peso percentuale dell'assessment sul voto finale.

date_submitted Giorno in cui lo studente ha inviato l'assessment.

is_banked Indica se il voto è stato "bancato" da un modulo precedente (Y/N).

score Punteggio ottenuto dallo studente nell'assessment (su 100).

3.2 La Task

L'analisi svolta in questo progetto prevede un task di classificazione su tre classi: l'obbiettivo è utilizzare i dati a disposizione per prevedere se uno studente passerà, non passerà o si ritirerà da un corso. La variabile che sarà utilizzata come target è `final_result`: originariamente le etichette previste erano quattro, ma si è deciso di aggregare la classe *distinction* all'interno di *pass*, così da evitare la presenza di classi troppo simili.

3.3 Data Understanding & Preparation

In questa fase preliminare, le variabili del dataset sono state esplorate al fine di migliorarne la comprensione, sono stati gestiti i valori mancanti, sono state rimosse informazini ridondanti o non utili all'analisi e svolto operazioni generali di pulizia dei dati. In questa fase, il dataset presenta oltre 170'000 istanze poiché sono ancora distinti tutti gli assignments per corso, dunque le analisi distribuzionali proposte sono sulla base di questi e non per `id_student`. Si segnala inoltre che non è stata effettuata la fase di outlier removal: questa decisione deriva dal fatto che comportamenti anomali nelle performance e caratteristiche degli studenti sono tipiche dello scenario scolastico e rimuoverle avrebbe implicato perdita di informazione.

3.3.1 Data Transformation

Una prima esplorazione del dataset ha portato all'individuazione di numerosi **NaN**: questi sono valori *not a number*, considerati mancanti. Per ogni colonna che presentava valori di questo tipo sono state proposti approcci diversi. La prima colonna affrontata è stata `date_registration`, la cui percentuale di valori mancanti era pressocchè nulla, solo lo 0.004%. I valori nulli sono stati sostituiti con la media rispetto alla colonna `date_registration`. La feature successiva presentava un numero decisamente più alto di valori mancanti: questi erano in `date_unregistration` il 92.5%. L'alto valore di NaN è dovuto al fatto che gli studenti che hanno completato il corso non hanno una data inserita, mentre quelli che si sono ritirati presentano *withdrawl*. Questa informazione è già contenuta in due altre colonne, `final_result` e `score`, dunque si è proceduto con l'eliminazione della colonna. Proseguendo, la feature `imd_band` presentava il 4.42% di NaN, che sono stati sostituiti con la media per regione. Infine, le colonne `date` e `score` presentavano rispettivamente l'1.6% e lo 0.09% di valori mancanti. Nel primo caso essi sono dovuti al fatto che l'assessment in questione è un esame: l'articolo di presentazione del dataset ci dice che gli esami si tengono sempre nell'ultima settimana di corso, dunque il valore nullo è stato sostituito con la feature `module_presentation_length`, a rappresentare l'ultimo giorno di corsi. I valori NaN in `score` invece, corrispondono tutti a assessment di tipo *tutor marked assignement*, il che significa che possono essere considerati come mancata consegna ed essere sostituiti con 0.

L'analisi è proseguita individuando ulteriori necessità di preprocessing per alcune delle variabili nel dataset. Innanzitutto, la feature `highest_education` presentava cinque diversi possibili valori: 'A Level or Equivalent', 'Lower then A Level', 'HE Qualification', 'Post Graduate Qualification' e 'No Formal Equals'. Questi ultimi due valori erano rispettivamente presenti solo per l'1.1% e lo 0.8% degli studenti, quindi l'informazione apportata non era decisiva. Si è dunque deciso di accorparli a 'HE Qualification' e 'Lower than A Level'.
Successivamente, si è notato che anche la variabile `age_band` presentava una granularità non utile all'analisi, poiché divisa in tre fasce: '0-35', '35-55' e '55¡='. Quest'ultima classe in particolar modo, presentava una quantità irrisoria di records, si è dunque preferito ristrutturare la colonna in modo da avere solo due classi: '0-35' e '35+'.
La variabile `is_banked` presentava una bassissima deviazione nella sua distribuzione, dunque è stata eliminata poiché non aggiungeva informazione utile alla nostra ricerca.
`date_registration` è stata ulteriormente processata poiché presentava una forte granularità dalla quale era difficile estrarre informazione: ne è stata effettuata la discretizzazione in cinque categorie utilizzando i quintili. Le nuove classi create sono le seguenti: *Iscrizione fortemente anticipata*, *Iscrizione anticipata*, *Iscrizione congruente* (all'inizio del corso), *Iscrizione tardiva*, *Iscrizione fortemente tardiva*, e sono state inserite sotto una nuova variabile chiamata semplicemente `registrazione`. La colonna `date_registration` è poi stata eliminata.
Successivamente, si è notato che anche la variabile `total_clicks` era fortemente granularizzata, dunque si è proceduto con la sua discretizzazione sulla base di quintili creando una nuova variabile `engagement` composta dalle seguenti classi: 'very_low', 'low', 'average', 'high', 'very_high'.
Infine, la feature `date_submitted` era anch'essa estremamente granulare, si è proceduto dunque alla sua discretizzazione in due classi: 'submission_on_time' e 'late_submission', identificate rispetto alla feature `date`, giorno della consegna dell'assessment. `date` e `date_submitted` sono poi state eliminate.

3.3.2 Analisi Distribuzionale e Statistiche Interessanti

Si presentano le analisi eseguite sulle features più rilevanti per l'obbiettivo dello studio, così come rilevazioni interessanti che sono state scoperte nella fase di studio dei dati.

La variabile target `final_result` si presenta con una distribuzione molto sbilanciata a favore dell'attributo *pass*, che è presente per un totale di 106'011 istanze, mentre gli altri valori sono presenti in quantità decisamente più basse. La classe meno popolosa è *withdrawn*, per un totale di 13'036 elementi, preceduta da *distinction*, con 26'330 istanze e *fail*, con 28367 records. Questo implica che aggregando *withdrawn* e *fail*, 39'366 assignments non sono stati passati, ovvero circa un quarto del totale.
Si è proceduto con l'analisi di alcune metriche aggregate per investigare eventuali connessioni tra variabili che potessero rappresentare conoscenza utile.

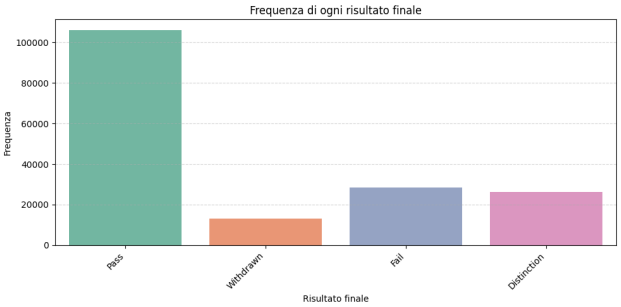

Figure 2: Distribuzione di `final_result`

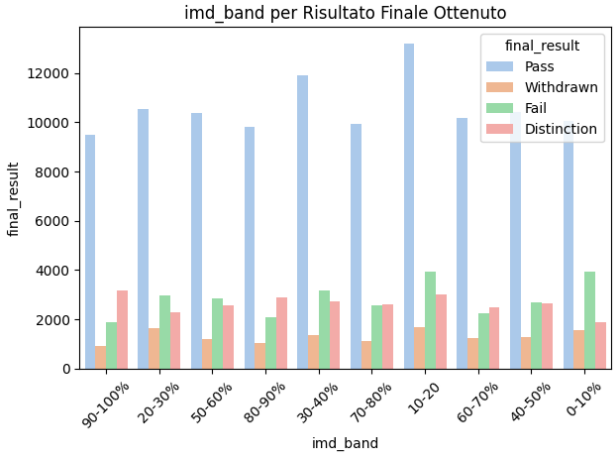

Figure 3: IMD Band per risultato finale

Rispetto alla variabile `gender`, il dataset è uniforme, con un 54% di studenti maschi ed un 46% di studentesse femmine. Osservando poi i loro risultati, non si notano differenze rilevanti. Al contrario, osservando la relazione tra la variabile `imd_band` e `final_result`, notiamo chiaramente come studenti provenienti da aree disagiate sotto al valore del 20% abbiano un tasso più alto di *fail* e *withdrawn*, come possiamo osservare nel grafico 3. Lo stesso vale per gli studenti con disabilità, che hanno rate più alti rispetto agli studenti senza disabilità per entrambe queste categorie [1]:

Table 1: Rate per ogni categoria di final result rispetto a studenti con o senza disabilità

final_result	**Studenti con disabilità**	**Studenti senza disabilità**
Pass	56.7%	61.4%
Fail	19.2%	16.1%
Withdrawn	11%	7.2%
Distinction	13.1%	15.4%

Analizzando invece le differenze per quanto riguarda le fasce di età, è apparso evidente come gli studenti appartenenti alla fascia 35+ ottenessero risultati più brillanti rispetto agli studenti nella fascia 0-35. Diminuisce il numero di istanze nella categoria *Pass*: nella fascia 0-35 è il 61.2% degli studenti a passare, mentre nella fascia 35+ è il 60.7%, ma diminuisce anche il numero di studenti che non hanno passato l'assessment (17.6% versus 13.6%), mentre aumenta il numero di promossi con `distinction` di ben 5 punti, 13.6% tra gli studenti con età inferiore ai 35 anni e 18.7% tra gli studenti con età superiore ai 35 anni, confermando che un'età più matura implica una consapevolezza diversa nello studio e migliore capacità di auto gestione.

Si è poi notato che la distribuzione della variabile `weight` rispetto al tipo di assessment presentava valori nettamenti più alti per l'assessment type *exam*, una piccola parte di teacher marked assessment aveva un

peso del 40%, mentre i computer marked assessment avevano peso zero per il voto finale.

L'analisi è proseguita con la visualizzazione delle distribuzioni delle variabili numeriche. Trend interessante è stato notato per la variabile `total_clicks`, che presenta una chiara distribuzione Zipfiana: pochissimi studenti hanno un conteggio di click molto alto, numero che scende seguendo il funzionamento di una Legge di Potenza. * [27] La distribuzione della variabile `date` invece, è bimodale: questo implica che la maggior parte degli assessment tende ad essere prevista in punti precisi del corso.

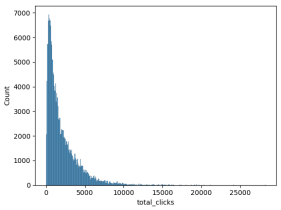

Figure 4: Distribuzione del numero di clicks

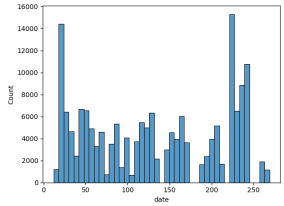

Figure 5: Distribuzione della data degli assessment

Infine, osservando la distribuzione della variabile `score`, si è notato un fatto sorprendente: la variabile presentava tre picchi, attorno a 60, 80 ed a 100. La classe di voto maggiormente popolate era, appunto, 100.

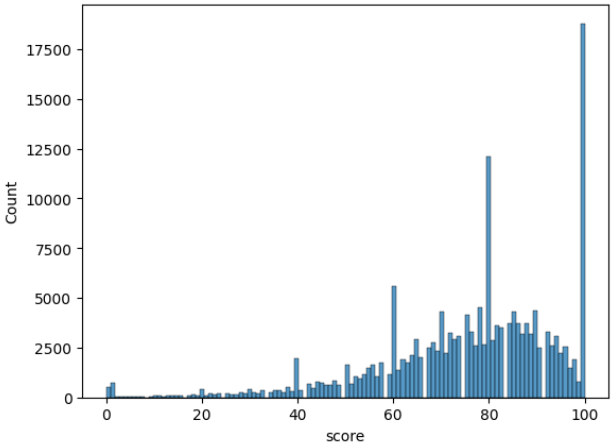

Figure 6: Distribuzione di `score`

3.3.3 Analisi della correlazione

L'analisi della correlazione coinvolge unicamente features numeriche, dunque per continuare ad approfondire come le caratteristiche degli studenti influenzino il loro successo sono state studiate principalmente due variabili: `total_clicks` e `date_submission` rispetto alla feature `score`. Il coefficiente di correlazione utilizzato è il **Coefficiente di Pearson**, che misura la forza e la direzione di una relazione lineare tra due variabili.

Dalla prima parte dello studio è emersa una correlazione positiva tra il numero di click totali effettuati dallo studente e lo score registrato per l'assessment, risultante al 20%: questo dimostra la legittimità di utilizzare un fattore come l'engagement calcolato tramite i click per quantificare l'impegno di uno studente nello studio. Nello scatterplot 7, possiamo osservare la tendenza in ascensione del rapporto tra le due features.

Per la seconda parte dello studio si è voluto utilizzare come indicatore di scarso engagement il numero di giorni in ritardo rispetto alla consegna di un assessment. Il valore è stato calcolato come la differenza tra la data di submission e la data dell'assessment e ne è stata calcolata la correlazione con la variabile `score`: la relazione è stata valutata come debolmente negativa poiché si situava approssimativamente attorno al -0.17%.

*Una legge di potenza (power law) è una qualsiasi relazione [...] dove a e k sono costanti e [la relazione tra loro] è una funzione asintoticamente piccola di x elevato alla k. k è di solito chiamato esponente di scala.

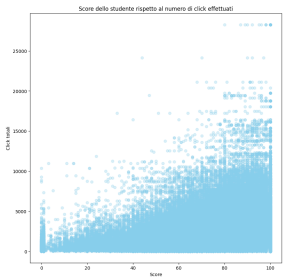

Figure 7: Correlazione tra total_clicks e score

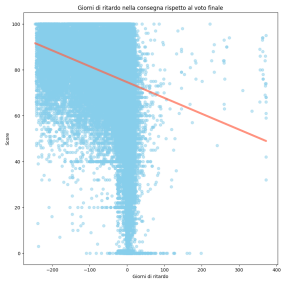

Figure 8: Correlazione tra date_submission e score

Altri rapporti interessanti che l'analisi di correlazione ha individuato coinvolgono sempre la variabile `total_clicks`. Sembra esserci una, seppure leggera, correlazione tra essa e `date` e `studied_credits` [Immagine 10]. Si nota come gli studenti con il numero più alto di clicks si iscrivino in anticipo ai corsi, ma entro i 150 giorni prima. Questo ci fa comprendere che gli studenti iscritti tra 1 e 5 mesi prima sono i più attivi: è possibile che l'iscrizione anticipata sia un indicatore di engagement. Nel plot a destra invece notiamo come studenti che hanno un numero di click maggiore tendono ad aver raggiunto un numero minore di crediti: è possibile che man mano che il percorso di studi procede, lo studente comprenda come ottimizzare il suo studio.

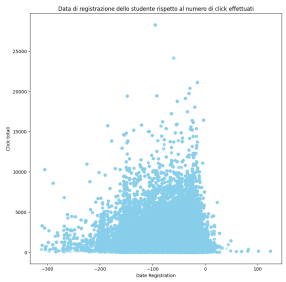

Figure 9: Correlazione tra total_clicks e date

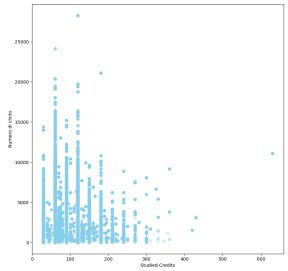

Figure 10: Correlazione tra total_clicks e studied_credits

4 Clustering

Le features del dataset risultante dalla fase di preparazione sono state ulteriormente processate in preparazione al clustering. Innanzitutto, sono state eliminate le features che non apportavano informazioni aggiuntive riguardo il comportamento del singolo studente o riguardo aspetti peculiari dei corsi che potessero avere conseguenze sulle performance. Le features `id_student`, `code_module` e `code_presentation` sono dunque state eliminate.
Le variabili numeriche sono poi state scalate utilizzando il `MinMaxScaler` della libreria Scikit-learn, il quale trasforma tutte le features in una scala tra 0 ed 1. Le istanze categoriche invece, sono state trasformate in vettori One-Hot, ovvero vettori che presentano ogni valore uguale a 0, fatta eccezione per la posizione della feature che si sta codificando, dove sono uguali ad 1.
Il clustering è stato effettuato utilizzando l'algoritmo `KMeans`: questo algoritmo "clusterizza i dati provando a separare esempi in N gruppi di varianza uguale, minimizzando un criterio conosciuto come inertia o within-cluster sum-of-squares"[15]. L'inertia misura quanto un cluster è coeso al suo interno, non è però una misura ottimale da utilizzare da sola, per due motivi principali: il primo è che non si ha una soglia di riferimento come valore minimo da rispettare, per quanto più il valore sia basso migliore sia il risultato. Inoltre, poiché utilizza la distanza euclidea per misurare le distanze tra due punti, non è ottimale se abbiamo dati ad alta dimensionalità, dato che le distanze in spazi ad alte dimensioni tendono a diventare simili. L'inertia è dunque molto spesso utilizzata in combinazione con la **silhouette**, coefficiente che misura quanto un punto è simile

agli altri all'interno del suo cluster, rispetto ai punti all'interno degli altri cluster. Maggiormente è vicina ad uno, maggiore ogni punto nel cluster è più simile ai suoi vicini.
Aspetto fondamentale del clustering con KMeans è la scelta di N (o K). Questo iperparametro deve essere scelto dall'utente: più sarà alto, più alto sarà il rischio di un eccessivo adattamento ai dati e al rumore, il che significa perdita di capacità di generalizzazione su nuovi dati da parte dell'algoritmo. Al contrario, più K sarà basso, meno sarà probabile trovare relazioni significative all'interno dei dati. Allo scopo di ovviare a questa problematica, un metodo comune per la scelta del K migliore è provare configurazioni diverse in più iterazioni, calcolando ogni volta l'inertia e la silhouette. Questo metodo è noto come **Metodo del Gomito**, poiché, tracciando i valori dell'inertia su un grafico in funzione di K, il numero ottimale di cluster corrisponde al punto in cui la curva forma un "gomito", ovvero il punto in cui la riduzione dell'inertia inizia a diminuire significativamente. Questa considerazione non è però una regola: per scegliere il miglior valore di K è fondamentale guardare anche al grafico della silhouette, scegliendo quello corrispondente al valore più alto della metrica
Infine, poiché i dati a disposizione si presentavano ad alta dimensionalità, per la visualizzazione dei cluster è stata usata la **Principle Component Analysis**[†][28].

I range di K testati in questo studio sono stati tre: il primo caso ha previsto l'analisi di K da 2 a 50, trovando il valore ottimale su 15. La divisione in quindici cluster non si è però rivelata significativa, poiché essi risultavano fortemente spuri e scarsamente separati. Si è dunque provato ad abbassare la soglia massima di K a 20, trovando questa volta $K = 19$ come valore ottimale, che però nuovamente risultava nella visualizzazione di cluster estremamente confusi. Poiché un numero elevato di cluster portava a confusione nell'interpretazione dei dati, l'utlimo caso di studio ha previsto un limite di K uguale a 6. Quest'ultima configurazione ha dato i risultati migliori. Come si può osservare in figura 11, il 'gomito' è riscontrabile in $K = 4$, seppure la silhouette più alta corrisponda a $K = 6$. La visualizzazione ha permesso di notare come i cluster trovati corrispondessero a relazioni significative poiché coerenti e separati. In figura 13, la rappresentazione dei cluster con l'utilizzo della PCA.

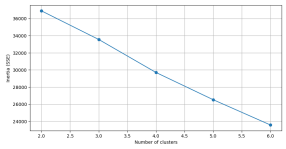

Figure 11: Inertia

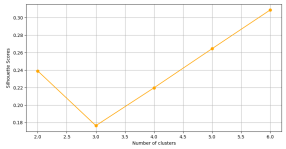

Figure 12: Silhouette

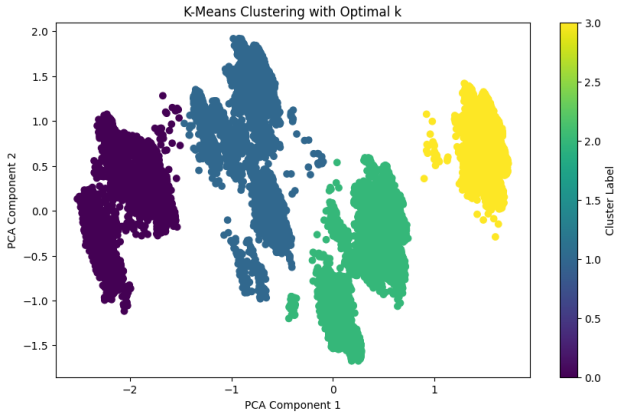

Figure 13: Visualizzazione dei cluster tramite PCA

Si è poi utilizzato un `Decision Tree Classifier` per analizzare quali fossero le features più impor-

[†]Tecnica per la semplificazione dei dati, di riduzione della dimensionalità lineare utilizzata nell'ambito della statistica multivariata.

tanti per la divisione in cluster degli studenti. È emerso che tra le prime cinque features solamente due riguardavano l'engagement dello studente: *late_submission* e *submission_on_time*. Le altre erano *gender_M*, *assessment_type_CMA* e *code_module_FFF*. Le features rimanenti erano presenti con influenza meno rilevante.

5 Classificazione e Feature Importance

Nella fase finale dello studio sono stati implementati alcuni modelli di machine learning con l'obiettivo di individuare le caratteristiche più influenti nell'andamento della carriera dello studente, così da indirizzare gli sforzi educativi verso ambiti specifici.
L'analisi si è divisa in due fasi: la prima parte ha previsto il training e la validazione di una serie di modelli per comprendere quali fossero i più adatti al dataset a disposizione. Tra questi sono poi stati scelti i due modelli con le performance migliori, per i quali è stata fatta un'approfondita ricerca nello spazio degli iperparametri così di aumentarne ulteriormente le prestazioni.
Il dataset preprocessato è stato suddiviso in 70% di dati utilizzati per il training e 30% di dati per il testing dei modelli. La variabile target è stata identificata in `final_result`, che si proietta su quattro classi: *pass*, *fail*, *distinction* e *withdrawn*. La classe *distinction* rappresentava una eccessiva granularità per lo scopo di questo studio, è dunque stata accorpata a *pass*. La feature target è stata codificata mediante binarizzazione, con l'utilizzo dell'oggetto `Label Binarizer`.

5.1 Training e Validazione

Il primo step dell'analisi ha previsto la scelta dei modelli da allenare e testare in fase di validazione. I modelli scelti sono stati i seguenti[21]:

- Support Vector Machine (SVM): applicato utilizzando il kernel radial basis function.
- Decision Tree Classifier (DT): metodo predittivo che utilizza una struttura ad albero per prendere decisioni, basandosi sugli 'split', decisioni *si* vs *no* prese sulla base degli attributi ritenuti più interessanti.
- Gradient Boosting Classifier: metodo predittivo ensemble, è una strategia che combina più modelli base (detti base learners o weak learners) per ottenere prestazioni predittive migliori rispetto a un singolo modello.
- Random Forest: altro metodo ensamble molto simile al Gradient Boosting, che utilizza come base learners i decision trees.
- K-Neighbors (KNN): metodo predittivo che classifica sulla base dei vicini dell'oggetto considerato.
- Logistic Regression: metodo che modella le log-odds di un evento basandosi sulla combinazione di una serie di features.
- Gaussian Naive Bayes: classificatore che presuppone una distribuzione di tipo gaussiano per ogni feature ed ogni classe, in questo studio è usato come baseline di paragone per gli altri modelli.[10]

Per effettuare la fase di validazione, il training set è stato ulteriormente diviso in due porzioni: l'80% di esso è rimasto tale, mentre il 20% è diventato il validation set. In tabella 2 sono inseriti i risultati della fase di validazione:

Gli algoritmi che hanno performato meglio sono stati i due modelli ensemble: Gradient Boosting e Random Forest. Quest'ultimo, insieme al Decision Tree, presenta un fortissimo overfitting sul training set, ovvero un completo adattamento ad essi con 0% di errore sulla predizione. Questo fenomeno era previsto dato che non sono state applicate in questa fase preliminare tecniche di prevenzione come, ad esempio, l'impostazione di una profondità massima per lo sviluppo degli alberi. Nonostante l'evidente overfitting sul training set, i modelli hanno comunque mostrato una buona capacità di generalizzazione sul set di validazione. In particolare, questo comportamento era atteso nel caso della Random Forest, poiché l'overfitting riguarda i singoli alberi costituenti, ciascuno addestrato su un sottoinsieme diverso dei dati. La combinazione aggregata delle

Modello	Accuracy sul Training Set	Accuracy sul Validation Set	F1 Score sul Training Set	F1 Score sul Validation Set
GradientBoost	0.81	0.79	0.80	0.78
Random Forest	1.00	0.77	1.00	0.75
Decision Tree	1.00	0.73	1.00	0.73
Logistic Regression	0.70	0.70	0.68	0.67
SVM	0.72	0.69	0.67	0.64
KNN	0.75	0.64	0.73	0.61
GaussianNB	0.47	0.47	0.47	0.47

Table 2: Performance dei modelli di classificazione

loro predizioni, tipica di questo approccio ensemble, consente di ridurre la varianza e preservare una buona generalizzazione.
In generale, tutti i modelli superano la performance della baseline GaussianNB, che ritorna il 47% di accuratezza, dunque ogni modello riesce ad imparare dai dati che gli sono messi a disposizione.

5.2 Testing

La fase di testing ha previsto la valutazione dei primi due modelli per performance: GradientBoosting e Random Forest. Al fine di migliorarne le performance, per entrambi è stata eseguita una cross-validation con ricerca randomica nello spazio degli iperparametri così da trovarne i valori più ottimali per la task da risolvere.
Il primo modello testato è stato Gradient Boosting. Gli iperparametri ottimizzati tramite `RandomizedSearch` sono stati i seguenti:

`learning_rate`	{0.01, 0.1, **0.2**}
`n_estimators`	{10, **100**, 1000}
`min_samples_split`	{**2**, 5, 10}
`min_samples_leaf`	{2, 3, 5, **10**}
`max_depth`	{None, 3, **5**, 10}
`ccp_alpha`	{**0**, 1}

In grassetto, i valori che sono stati scelti. Per quanto riguarda il Random Forest Classifier invece, si riportano in seguito gli iperparametri considerati nel processo di ottimizzazione del modello. Anche in questa tabella i valori prescelti sono segnati in grassetto:

`n_estimators`	{10, **100**} (per Random Forest)
`max_depth`	{10, 20, 30, **None**} (Random Forest)
`min_samples_split`	{2, 5, **10**}
`min_samples_leaf`	{1, **2**, 4}
`class_weight`	{**None**, "balanced"}

Infine, riportiamo i risultati per entrambi i modelli sul test set.

Poichè il dataset di partenza era molto sbilanciato in favore di *Pass*, era atteso che questa fosse la classe più facilmente individuabile, nonostante i classificatore abbiano ricevuto esplicita indicazione di considerarle tutte con lo stesso peso.

(a) Risultati del Gradient Boosting Classifier

Classe	Precision	Recall	F1-Score
Fail	0.56	0.49	0.52
Pass	0.89	0.95	0.92
Withdrawn	0.62	0.53	0.57
Accuracy totale	0.78		

(b) Risultati del Random Forest

Classe	Precision	Recall	F1-Score
Fail	0.54	0.45	0.49
Pass	0.84	0.97	0.90
Withdrawn	0.61	0.38	0.47
Accuracy totale	0.76		

Table 3: Risultati delle fasi di testing

5.3 Feature Importances

La parte finale di questo studio si concentra sull'analisi delle features più importanti usate per la classificazione di ogni classe. Lo scopo di questo approfondimento consiste nell'individuare gli aspetti nella carriera dello studente che maggiormente incidono sul suo successo.

Cominciando con `GradientBoosting`, le variabili maggiormente informative per ogni classe sono state identificate accedendo all'elemento `feature_importances` e sono esposte in tabella 4:

Table 4: Attributi principali per l'individuazione della classe *Fail*

Feature	**Importance**
weight	0.38
score	0.20
total_clicks_very_low	0.18
assessment_type_TMA	0.04
module_presentation_length	0.04
assessment_type_CMA	0.02
studied_credits	0.02
submission_submission_on_time	0.00846
submission_late_submission	0.007957
num_of_previous_attempts	0.006

Le prime due variabili sono chiaramente molto informative riguardo l'andamento della carriera dello studente. Il primo attributo è il peso dell'assessment in questione: prove con un peso maggiore implicano maggiore difficoltà e quindi probabilità più alta di non passare. Secondo è il voto che lo studente ha ricevuto nel corso. Un valore inferiore a 40 è considerato una bocciatura, dunque era atteso che il modello guardasse specialmente ad esso. Successivamente, si nota come il modello abbia utilizzato una feature di engagement molto significativa, il numero di clicks effettuati: se questa presenta un valore basso, è molto probabile che lo studente non sia in grado di passare l'esame. Si nota poi la presenza dei due tipi di assessment con peso minore: questi influenzato meno il voto finale ma la loro considerazione potrebbe portare alla luce trend di comportamento durante tutto il corso. Fondamentale è anche `module_presentation_length`: corsi con durata maggiore implicano maggiore mole di studio e maggiore difficoltà. Si possono poi osservare altre due variabili interessanti: il numero di crediti conseguiti e, soprattutto, entrambe le variabili relative alla consegna dell'assessment. Come ci si aspettava, la consegna per tempo implica un'organizzazione migliore da parte dello studente e, dunque, performance migliori. Infine, il numero di tentativi precedenti a quello in esame. Il modello ha compreso che questa variabile potrebbe rappresentare un pattern: fallimenti precedenti possono implicare fallimenti futuri.
Si presentano adesso, in tabella 5, le prime dieci features per la classificazione effettuata dal `RandomForest`.

Si rileva una parziale sovrapposizione tra le feature inizialmente selezionate dai modelli, evidenziando la presenza di un pattern informativo condiviso e potenzialmente rilevante. In aggiunta alle variabili individuate nello step precedente, il modello `RandomForest` ha selezionato anche la variabile *cluster_label*, suggerendo

Table 5: Attributi principali per l'individuazione della classe *Fail*

Feature	**Importance**
weight	0.26
score	0.17
total_clicks_very_low	0.08
module_presentation_length	0.05
cluster_label	0.05
studied_credits	0.03
assessment_type_TMA	0.02
assessment_type_CMA	0.02
total_clicks_very_high	0.01
submission_late_submission	0.01

che i cluster ottenuti mediante `KMeans` siano effettivamente significativi. Inoltre, la presenza della variabile *total_clicks_very_high* tra le feature rilevanti indica che un livello particolarmente elevato di engagement risulta associato in modo significativo alla direzione del percorso accademico degli studenti.

6 Conclusioni

L'analisi ha evidenziato quali siano gli aspetti più importanti nella carriera di uno studente, portando a conclusioni importanti che possono essere potenzialmente sfruttate in ambito educativo per rendere il percorso di istruzione maggiormente proficuo.
Dalle variabili emerse durante lo studio dei modelli, è chiaro innanzitutto come informazioni di engagement debbano essere considerate di primaria importanza. In un percorso educativo è necessario dunque riuscire a proporre materiali utili ed interessanti così da poter coinvolgere lo studente nel loro completamento, ma anche assicurarsi che egli sia seguito nella realizzazione delle prove, così da completarle in un tempo utile. Fondamentale risulta essere anche il sostegno per quanto riguarda i corsi impegnativi: proponendo eventualmente lezioni solo di chiarimento o gruppi di studio che permettano di affrontare in modo proficuo anche altri tipi di prove che non siano l'esame finale. Infine, risulta di grande importanza porre un interesse particolare sugli studenti che hanno numero di tentativi precedenti per esame più alto della media, di nuovo proponendo supporto in varie forme: dall'aiuto allo studio alla comprensione verso le motivazioni che hanno portato a risultati non positivi.
In conclusione, l'analisi condotta ha permesso di identificare un insieme coerente di variabili predittive associate all'insuccesso degli studenti. Le feature più rilevanti includono indicatori di scarso engagement, performance intermedie insufficienti e una bassa interazione con le risorse didattiche digitali. Tali risultati suggeriscono la possibilità di intervenire precocemente attraverso strategie mirate di supporto, orientamento e personalizzazione dell'apprendimento. Studi futuri potranno approfondire la generalizzabilità di questi pattern su altri contesti didattici e valutare l'efficacia di interventi preventivi basati su questi segnali predittivi.

References

[1] A. N. Amorim, L. Jeon, Y. Abel, E. F. Felisberto, L. N. Barbosa, and N. M. Dias. Using escribo play video games to improve phonological awareness, early reading, and writing in preschool. *Educational Researcher*, 49(3):188–197, 2020.

[2] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23:537–553, 2018.

[3] P. Brusilovsky. Adaptive hypermedia: From intelligent tutoring systems to web-based education. In *International Conference on Intelligent Tutoring Systems*, pages 1–7. Springer, 2000.

[4] dataschop. https://pslcdatashop.web.cmu.edu/index.jsp?datasets=public.

[5] Duolingo. `https://sharedtask.duolingo.com/2018.html`. Ultimo accesso: 7 luglio 2025.

[6] R. K. Ellis. Learning management systems. *Alexandria, VI: American Society for Training & Development (ASTD)*, 2009.

[7] R. Ferguson. Learning analytics: fattori trainanti, sviluppi e storie. *Italian Journal of Educational Technology*, 22(3):138–147, 2014.

[8] A. C. Graesser, M. W. Conley, and A. Olney. Intelligent tutoring systems. *American Psychological Association*, 2012.

[9] Kaggle. `https://www.kaggle.com/datasets/sst2023/kdd-cup-2015`. Ultimo accesso: 7 luglio 2025.

[10] Kaggle. https://www.kaggle.com/code/yousefalbasel/oulad-personalized-learning-path-recommender-sys/notebook. Ultimo accesso: 15 luglio 2025.

[11] kddcup. https://pslcdatashop.web.cmu.edu/KDDCup/.

[12] kddcup. `/biendata.com/competition/kddcup2015/`. Ultimo accesso: 7 luglio 2025.

[13] A. Kirkwood and L. P. and. Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? a critical literature review. *Learning, Media and Technology*, 39(1):6–36, 2014.

[14] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.

[15] S. learn library. https://scikit-learn.org/stable/modules/clustering.html#k-means. Ultimo accesso: 14 luglio 2025.

[16] J. Luan. Data mining and knowledge management in higher education-potential applications. *ERIC*, 2002.

[17] M. C. Mihaescu and P. S. Popescu. Review on publicly available datasets for educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3):e1403, 2021.

[18] H. MITx. Harvardx-mitx person-course academic year 2013 de-identified dataset, version 2.0. harvard dataverse (2014).

[19] C. Romero and S. Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, 3(1):12–27, 2013.

[20] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355, 2020.

[21] Scikit-learn. https://scikit-learn.org/stable/index.html. Ultimo accesso: 15 luglio 2025.

[22] https://archive.ics.uci.edu/ml/index.php.

[23] https: https://data.mendeley.com/.

[24] https: https://dataverse.harvard.edu/.

[25] e. a. Vahdat. Educational Process Mining (EPM): A Learning Analytics Data Set. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5NP5K.

[26] Wikipedia. https://en.wikipedia.org/wiki/Open$_U niversity.Ultimoaccesso$: 15$luglio$2025.

[27] Wikipedia. https://it.wikipedia.org/wiki/Legge$_d i_p otenza. Ultimo accesso$: 14$luglio$2025.

[28] Wikipedia. https://it.wikipedia.org/wiki/Analisi$_d elle_c omponenti_p rincipali. Ultimo accesso$: 14$luglio$2025.

[29] J. Wulf, I. Blohm, J. M. Leimeister, and W. Brenner. Massive open online courses. *Business & Information Systems Engineering*, 6:111–114, 2014.