ANALISI PREDITTIVA SUL DATASET OULAD

Lavinia Rotellini

OVERVIEW

1 Presentazione
2 Introduzione
3 Realizzazione di un dataset in ambito EDM
4 Specificità del Pre-processing
5 Il dataset OULAD
6 Introduzione seconda parte
7 Analisi delle variabili
8 Clustering
9 Feature Importance e Conclusioni

INTRODUZIONE

La prima parte del lavoro presenta una breve **ricognizione sul TEL**, oltre a specificare le **metodologie di lavoro** con dataset EDM e le **risorse principali** in questo ambito.

La task proposta nella seconda parte dello studio è un'**analisi predittiva** di classificazione con **feature importance** per comprendere gli aspetti più significativi che determinano l'andamento della carriera di uno studente.

Lo studio presentato oggi si situa nell'ambito del **technology-enhanced learning**, campo di studi nato negli anni '90.

Il TEL si articola in **tre ambiti**:

- educational data mining
- learning analytics
- academic analysis

In particolare, il presente lavoro si situa nell'ambito **dell'educational data mining**: lo studio di dati di tipo educativo allo scopo di scoprire schemi e realizzare predizioni sui comportamenti degli studenti.

REALIZZAZIONE DI UN EDM DATASET

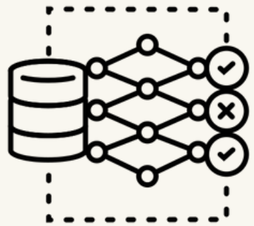
Task con molti fattori di complessità per la quale possiamo avere cinque diverse applicazioni:

Evaluations

Scientific Inquiry

Ma anche:

- Student Modeling
- Adaptive Systems



Decision Support Systems

I dati sono reperibili da fonti diverse:

Face-to-face learning

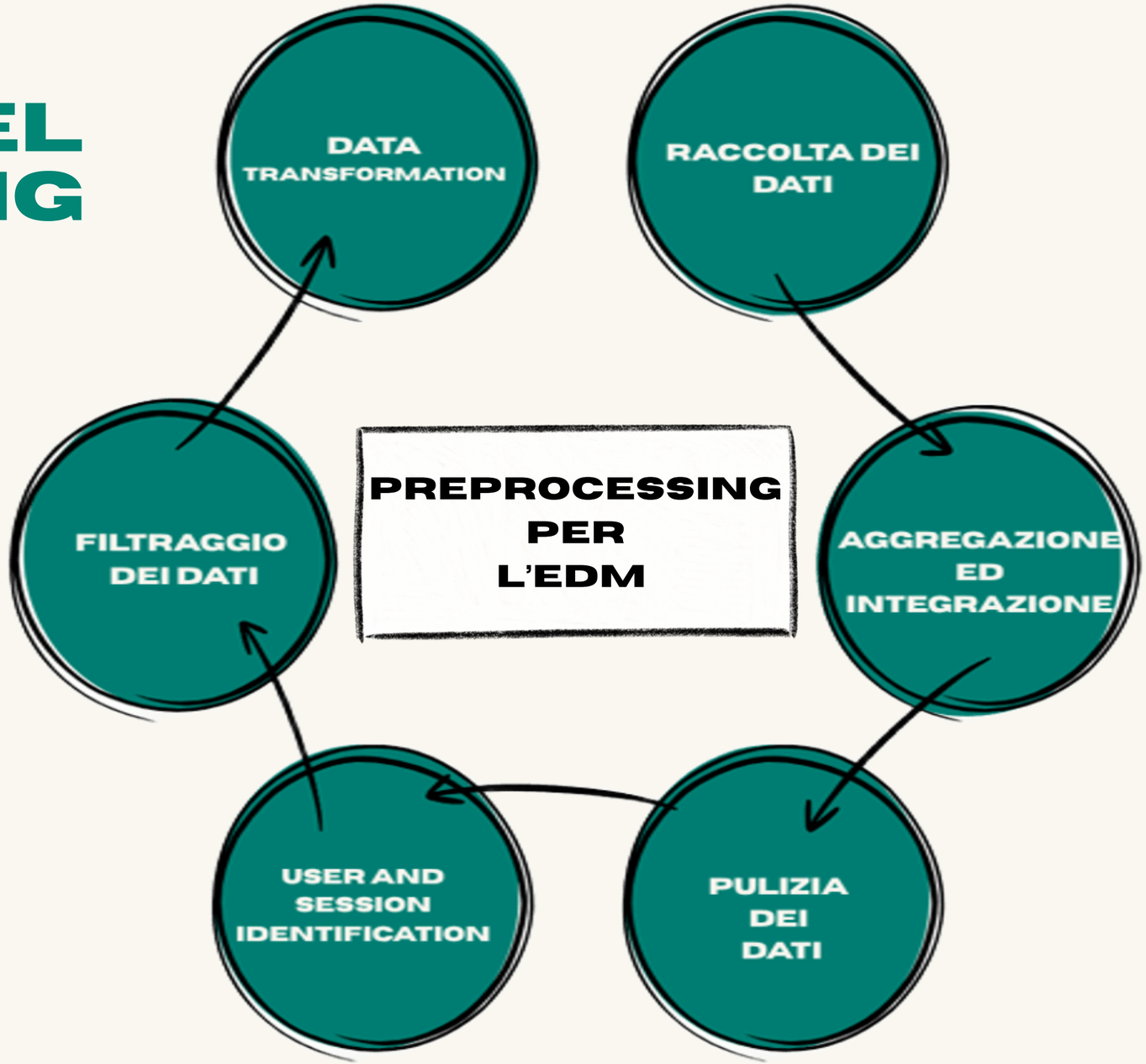
Virtual Learning Environments

Learning Management Systems

Massive Online Open Courses

SPECIFICITA' DEL PREPROCESSING

I dati nell'educational data mining presentano alcune peculiarità che implicano trattamenti specifici.



I DATASET PER L'EDM

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

DataShop @CMU

HARVARD Dataverse

MENDELEY DATA

OULAD DATASET

È stato scelto per **tre ragioni**:

- Affermato in letteratura
- Varietà di informazioni
- Accessibilità

Caratteristiche principali:

- Student-oriented
- Tre tipi di dato
- Organizzazione per moduli

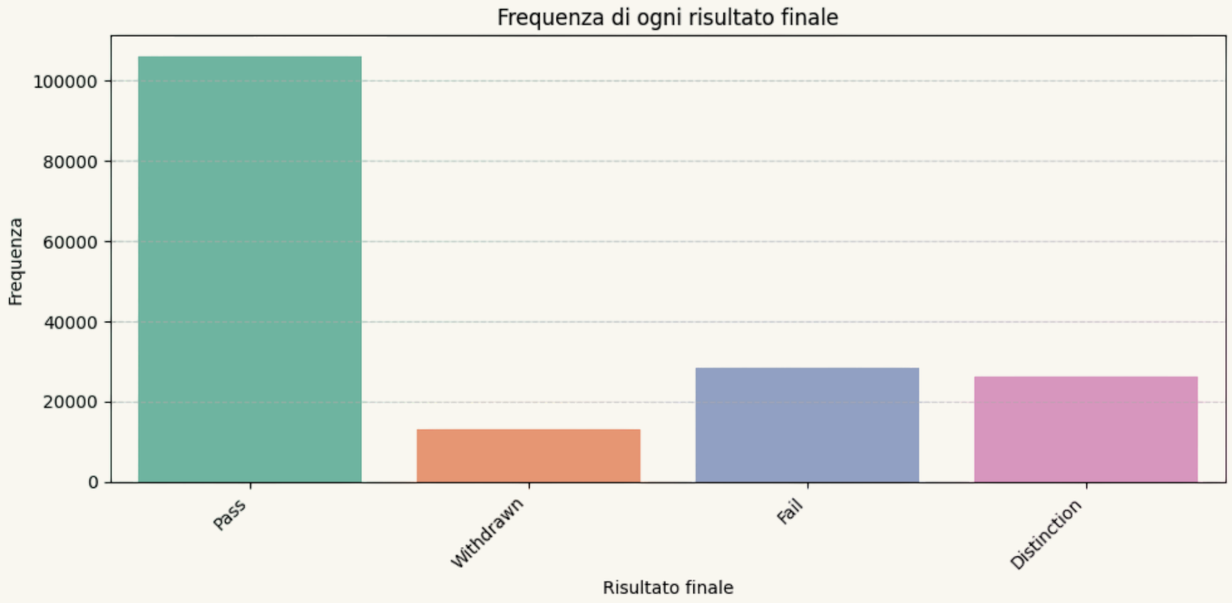
SECONDA PARTE: ANALISI PREDITTIVA E FEATURE IMPORTANCE

ANALISI DELLE VARIABILI

- Le tabelle sono state unite con un **inner join**
- Aggregate sulla base di **id_student**
- Specifiche di data transformation: **NaN** e **granularità** delle features
- Analisi della correlazione rispetto a `score`

Analisi distribuzionale di `final_result`

- Gender
- imd_band
- Disability

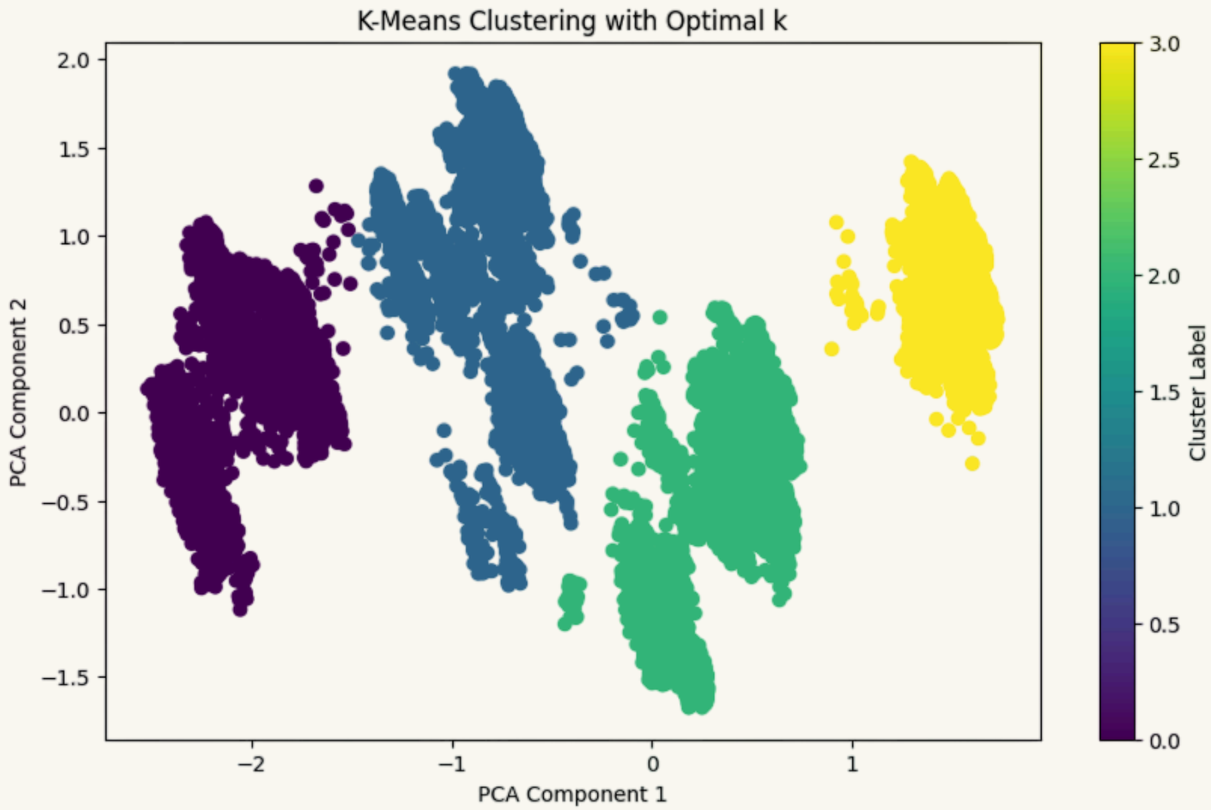


CLUSTERING

È stato effettuato il **clustering** dei dati con visualizzazione tramite **PCA e k =4**

I cluster identificati presentavano le seguenti caratteristiche:

- **0**: F, very_low, iscrizione fortemente tardiva, late_submission
- **1**: M, very_low, iscrizione_fortemente_anticipata, submission_on_time
- **2**: F, very_low, iscrizione_fortemente_anticipata, submission_on_time
- **3**: M, very_high, iscrizione fortemente tardiva, submission_on_time



FEATURE IMPORTANCE E CONCLUSIONI

L’analisi condotta ha permesso di identificare un insieme coerente di variabili predittive associate all’andamento della carriera di uno studente. Le feature più rilevanti individuate sono le seguenti:

- **Indicatori di engagement: total_clicks, submission**
- **Performance intermedie: assessment_type, weight**
- **Difficoltà inerente ai corsi: module_presentation_length**

Tali conoscenze possono essere implementate nella didattica sotto diversi aspetti:

- Proporre materiali utili ed interessanti così da poter coinvolgere lo studente nel loro completamento
- Assicurarsi che lo studente sia seguito nella realizzazione delle prove, così da completarle in un tempo utile
- Sostegno per quanto riguarda i corsi impegnativi
- Porre un interesse particolare sugli studenti che hanno numero di tentativi precedenti per esame più alto della media

Grazie per l’attenzione