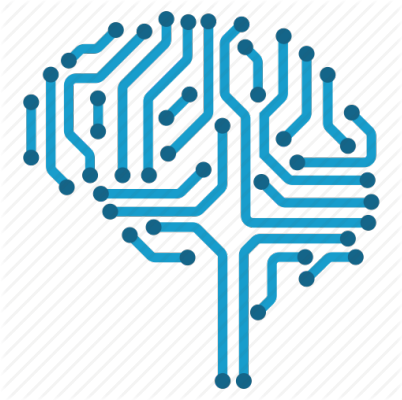
Big Data e Business Intelligence



Machine Learning

Giulio Angiani - UniPr

Machine Learning
Casi di Studio

Casi di Studio - ELDM

ELDM - Electronic Logbook Data Mining

- *Contesto*: Analisi dei dati degli studenti delle scuole italiane
- *Domanda*: E' possibile estrarre informazioni per anticipare situazioni di difficoltà prima possibile?
- *Domanda*: Quali dati sono utilizzati oggi?
  - INVALSI
  - PAGELLE
  - OECD - PISA
    - Solo dati già aggregati
- *Collaborazioni*: Regione Emilia Romagna - Università di Pisa - KDD Lab

ref: https://link.springer.com/chapter/10.1007/978-3-319-98872-6_5

Casi di Studio - ELDM

ELDM - Electronic Logbook Data Mining

- *Ricerca dei dati*: Recupero dati dei registri elettronici delle scuole superiori
- *Problemi*:
  - Anonimizzazione
  - Omogeneizzazione
  - Costruzione di features
- *Obiettivi*:
  - Mostrare le **best practices** educative esistenti
  - Evidenziare **patterns** comuni
  - Prevedere in anticipo situazioni **problematiche**

Casi di Studio - ELDM

Recupero dei dati

- 10 scuole italiane (Licei, Tecnici, Professionali, Sud, Centro, Nord)
- dati di 3 anni scolastici
- esportazione a carico delle società che fanno registri elettronici
  - dati anonimizzati
- formati esportati
  - csv
  - xml
- dati relativi a circa 13000 studenti
  - valutazioni
  - assenze
  - esiti di fine trimestre, fine anno, recuperi settembre

ELDM - data collect

Esempio di dati esportati

DATA : MATERIE

```
IDMATERIA,DESCRIZIONE,AREADISCIPLINARE
167741,BIOLOGIA, MICROBIOLOGIA E TECNOLOGIE DI CONTROLLO AMBIENTALE,
167735,BIOLOGIA, MICROBIOLOGIA E TECNOLOGIE DI CONTROLLO SANITARIO,
167727,CHIMICA ANALITICA E STRUMENTALE (SAN),
167726,CHIMICA ORGANICA E BIOCHIMICA (AMB),
```

DATA : VOTI

```
IDSTUDENTE,IDMATERIA,GIORNO,TIPODATO,VALORE
4054231,159174,2015/11/09,Scritto,4.000
4702763,159174,2015/11/09,Scritto,7.000
4702608,159174,2015/11/09,Scritto,7.000
4702807,159174,2015/11/09,Scritto,8.000
```

DATA: ASSENZE

```
IDSTUDENTE,GIORNO
622776,2015/09/18
622776,2015/09/19
```

ELDM - data collect

DATA : STUDENTI

```
IDSTUDENTE,ANNODICORSO,GRUPPODICLASSE,ANNOSCOLASTICO,
CODICEMECCANOGRAFICOSCUOLA,IDCORSO,SESSO,CODICESIDI,VOTOITALIANO1Q,
VOTOMATEMATICA1Q,NUMEROINSUFFICIENZE1Q, MEDIA1Q,VOTOITALIANO2Q,VOTOMATEMATICA2Q,
NUMEROINSUFFICIENZE2Q,MEDIA2Q,ESITOGIUGNO,ESITOSETTEMBRE
4702579,1,10,2015-2016,PRTF010006,93,M,6600261,3.00,3.00,10,3.580,,,0,,2,
1022142,4,2,2015-2016,PRTF010006,178,M,6045596,7.00,8.00,0,7.100,7.00,8.00,0,7.60,1,
4054281,2,7,2015-2016,PRTF010006,90,M,6045431,6.00,4.00,2,6.460,6.00,5.00,1,6.31,3,1
4229630,4,9,2015-2016,PRTF010006,176,M,7607278,6.00,6.00,0,6.700,6.00,6.00,0,6.90,1,
```

DATA : CORSI DI STUDIO

```
IDCORSO,CODICEMINISTERIALECORSO,DENOMINAZIONECORSO
93,IT13,INFOR. TELECOM. - BIENNIO COMUNE
178,ITMM,MECCANICA E MECCATRONICA
90,IT05,MECC. MECCATRON. ENER. - BIENNIO COMUNE
176,ITIA,INFORMATICA
172,ITET,ELETTROTECNICA
```

ELDM - Feature detection

- due tipi di valori
- granularità del dato:
  - lo studente S nel giorno D ottiene la valutazione V nella materia M
  - lo studente S è assente nel giorno D
- problema: **dati troppo grezzi**
  - individuazione istanze con valori *mancanti*
  - costruzione di valori *aggregati*
  - scelta di valori *significativi* per la ricerca
    - temporali
    - per disciplina/gruppi di discipline
  - *relativizzazione* del dato

ELDM - Ricerca

- selezione dei dati
- creazione features significative (medie ponderate, varianze, distanze)
- analisi per periodo/tipologia/discpline ogni mese 6 gruppi di materie (ita, ing, mat, sto, indirizzo, altro)

Machine Learning
Esercitazione

Esercitazione

- scaricare il dataset EDM da **http://www.giulioangiani.com/bdbi/edm_dataset.csv**
- Vi chiediamo di:
  - testare 2 classificatori diversi sul dataset
  - provare ad applicare tecnica di clustering
  - controllare quanti sono clusterizzati correttamente e quanti no
  - cercare le 20 features più significative
  - proiettare il dataset su tali features
  - testare i 2 classificatori precedenti sul dataset ridotto
  - controllare le differenze in termini di prestazioni e individuare eventuali elementi classificati diversamente da uno e dall'altro

Giulio Angiani
Universita' degli Studi di Parma