Business Intelligence per Big Data

Progetti di analisi di dati

AA 2025-2026 - *Politecnico di Torino*

Progetti di analisi di dati

- Attività da svolgere
  - Scegliere uno use case tra i due proposti
  - Caratterizzare il dataset
    - Esplorazione dei dati
  - Preprocessing opportunamente scelto per lo use case
  - Effettuare diverse sessioni di analisi sul dataset utilizzando il tool RapidMiner e/o Python e/o altri tools noti al team
    - Almeno 2/3 algoritmi
    - Analisi di sensitività dei parametri
  - Analizzare i risultati e sintetizzarli in grafici
    - Analisi comparativa

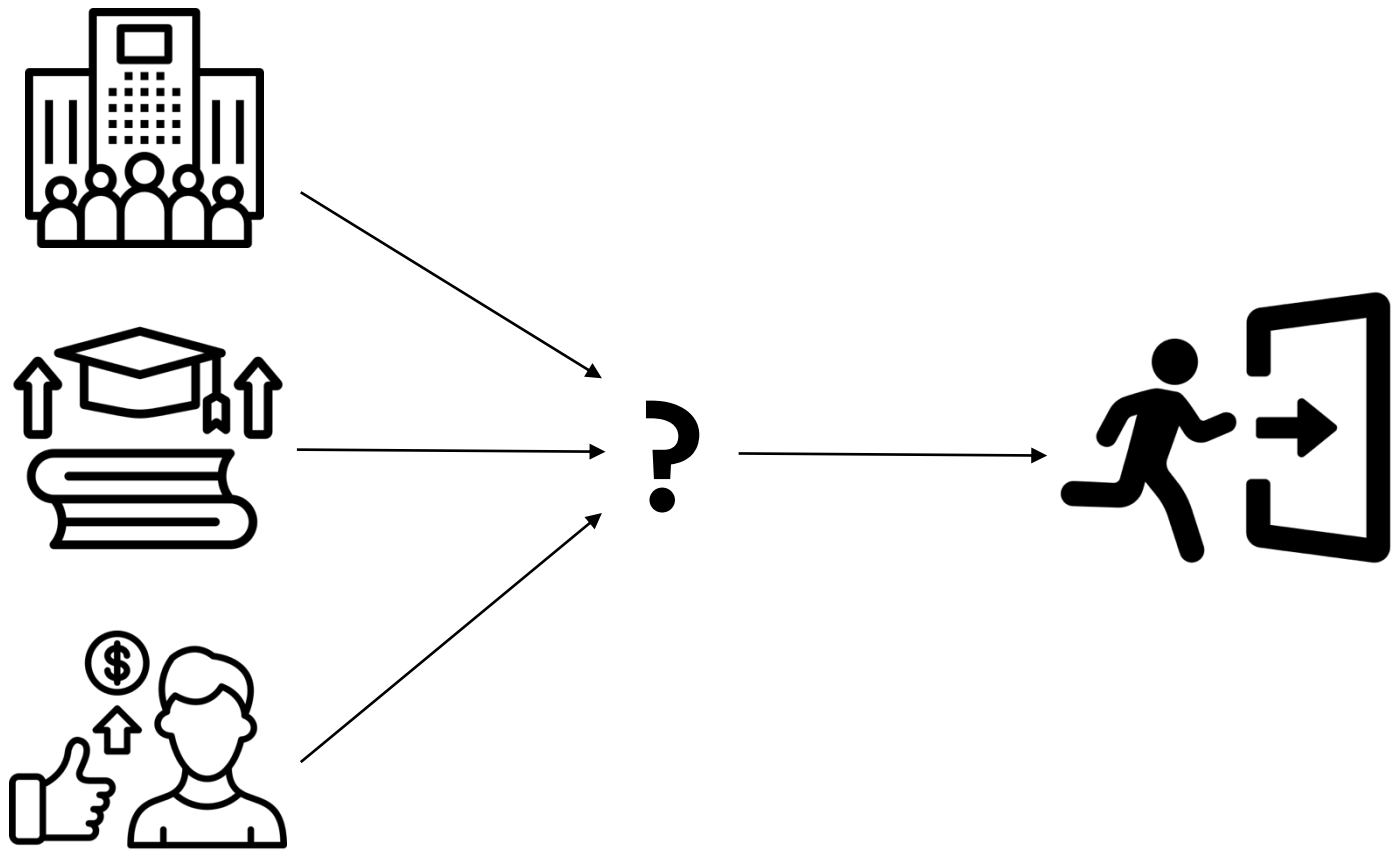
Progetti di analisi di dati

- Attività da svolgere (cont.)
  - Discutere come sfruttare la conoscenza estratta in un'applicazione di business
  - Scrivere il **report scientifico sintetico** sulle attività svolte
    - Utilizzare il template overleaf
    - Rispettare la struttura data
    - Rispettare il numero di pagine massimo (4 pagine massimo)

Formazione dei gruppi

- Formazione dei gruppi
  - Compilare il google form https://forms.gle/qyPDaqPYHWKCgepW6 entro 30 Aprile 2026 ore 23:59
  - 1 compilazione per gruppo
- Scelta dello use case
- Informazioni dei componenti del gruppo
  - Matricola, Nome, Cognome
- Le assegnazioni degli use case ai gruppi sarà resa disponibile entro il 6 Maggio 2026
  - Solo per variazioni per gruppi<5 persone e/o studenti e studentesse senza gruppo

Use Case: Classificazione dell'abbandono aziendale (attrition)

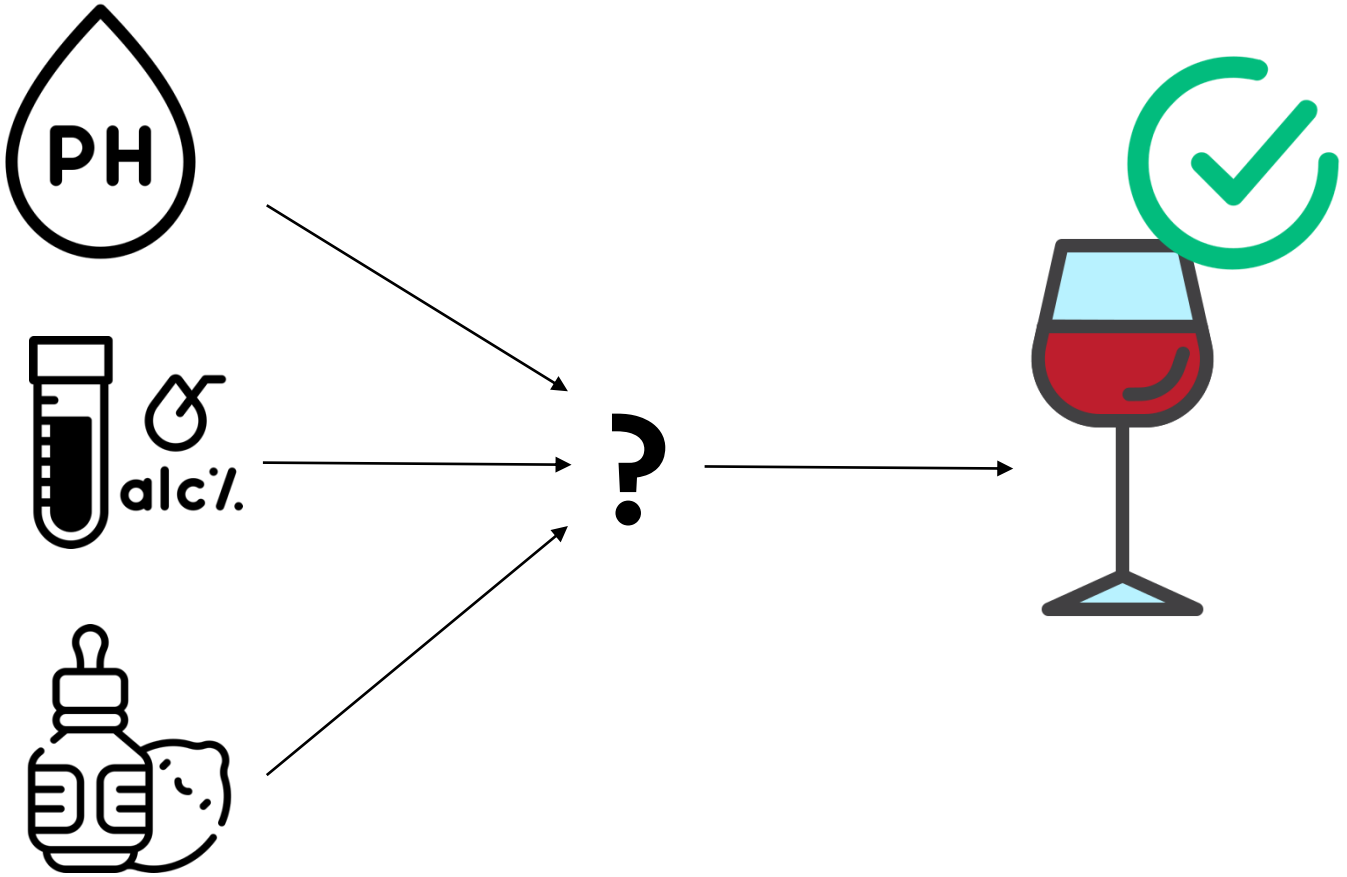


Caratteristiche del Problema

- L'attrition è un **costo nascosto elevato** (recruiting, formazione, produttività)
- Necessità di **prevedere in anticipo** i dipendenti a rischio di abbandono
- Problema di **classificazione binaria** (target: *Attrition* = 0/1)
- Dataset con **dati anagrafici e lavorativi** dei dipendenti
- Suddivisione in **train/test** (eventuale validation set)

Obiettivo: **supportare strategie di retention e decisioni HR**

Use Case: Classificazione della qualità del vino rosso



Caratteristiche del Problema

- Valutazione della qualità del vino tradizionalmente **costosa, soggettiva e poco riproducibile**
- Necessità di un sistema **automatico e standardizzato** basato su dati fisico-chimici
- Problema di **classificazione supervisionata multiclasse** (target: *WineQuality*, scala 3–8)
- Dataset strutturato (feature chimiche + etichetta qualità) diviso in **train/test**

Obiettivo: **ridurre costi, migliorare coerenza e supportare decisioni produttive**

Consulenze per le attività progettuali

- Slot di laboratorio
  - Martedì 26 Maggio 14:30-17:30
  - Mercoledì 3 Giugno 13:00-16:00
  - Slot da 1.5h da definire
- Mail
  - eleonora.poeta@polito.it
  - + eliana.pastor@polito.it

Materiale da consegnare

- Il gruppo deve consegnare
  - Report scientifico sintetico
  - Progetto overleaf: sorgenti latex
  - Processo di rapid miner e file memorizzati nel repository e/o codice
  - File excel/csv con i completi dati degli esperimenti (opzionale, risultati completi e/o aggiuntivi rispetto a quanto riportato nel report)

- Il gruppo deve compilare
  - 1-2 questionari online (obbligatori)

Date di consegna

- **Entro 7 giorni prima** della data della prova scritta
- Il gruppo deve consegnare un'unica cartella zip con il materiale indicato in slide 'Materiale da consegnare' effettuando l'upload sul Portale della Didattica, 'Sezione Elaborati' come descritto alla sezione 'Istruzioni per la consegna' della pagina del corso .

Valutazione

- Ogni studente del gruppo sarà valutato con un punteggio in trentesimi
  - Con valutazione pari 30 e lode, viene considerato 32 (per calcolare il voto finale)
    - Completezza, Metodo, Analisi Dati, Presentazione.
- Il voto del report sarà mediato con il voto conseguito all'esame scritto incrementato di
    - 1/30 in caso di consegna dell'homework su Google Data studio
    - 1/30 in caso di consegna dell'homework su MongoDB
- La lode viene riconosciuta se il voto non arrotondato è >=31