



UNIVERSITÀ  
DEGLI STUDI DI MILANO-BICOCCA

## COURSE SYLLABUS

### Data Mining

2122-3-E4101B026

---

#### Obiettivi formativi

Il corso intende fornire una visione completa del Data Mining, dal pre processamento del dato fino alla selezione del miglior modello statistico per l'analisi e la comprensione del problema.

Alla fine del corso, lo studente sarà in grado di confrontare e selezionare il miglior metodo di Data Mining per il problema oggetto di analisi. Saprà trattare le principali problematiche relative al dato e, autonomamente, affrontare un problema reale nel miglior modo.

#### Contenuti sintetici

Durante il corso verranno affrontate le principali tecniche per il trattamento dei dati e spiegati sia metodi statistici di tipo supervisionato sia non supervisionato. Inoltre verranno introdotti concetti relativi al Text Mining.

#### Programma esteso

1. Introduzione al Data mining. Concetti introduttivi e esempi applicativi
2. Pre-processing: trattamento dei missing values.
3. Introduzione alla classificazione con esempi e concetti introduttivi. Metodi di classificazione: regressione logistica, discriminante lineare, discriminante quadratico e k-nn.
4. Definizione di overfitting e tecniche per evitarlo
5. Introduzione al clustering con esempi e concetti introduttivi: metodi gerarchici e partizionali.
6. Text mining con esempi e concetti di base: pre-processing (stop words, stem words, ...), rappresentazioni grafiche e utilizzo del clustering per il Text Mining.

#### Prerequisiti

Analisi Statistica Multivariata e programmazione in R.

## **Metodi didattici**

Lezioni frontali e laboratorio.

## **Modalità di verifica dell'apprendimento**

Progetto e esame orale.

### **Scritto**

Prova scritta mirata a verificare gli argomenti presentati in classe.

### **Progetto**

Progetto applicativo da svolgere autonomamente o in gruppo su un dataset assegnato dal docente o scelto dagli studenti. Il progetto è realizzato in R e deve dimostrare la capacità di affrontare un problema reale in ogni suo aspetto utilizzando quanto visto a lezione.

Il progetto si compone sia del codice R sia di un report di presentazione.

### **Orale**

Presentazione e discussione del progetto.

Nel periodo di emergenza Covid-19 gli esami orali saranno solo telematici. Verranno svolti utilizzando la piattaforma WebEx e nella pagina e-learning dell'insegnamento verrà riportato un link pubblico per l'accesso all'esame di possibili spettatori virtuali.

### **Note**

Prove intermedie non sono previste.

Gli studenti non frequentanti sono pregati di contattare il docente almeno 15 giorni prima della data dell'esame.

## **Testi di riferimento**

Gareth J., Witten D., Hastie T., Tibshirani R., *An Introduction to statistical learning with application in R*, springer (2013).

Altro materiale verrà indicato a lezione.

## **Periodo di erogazione dell'insegnamento**

I Semestre

## **Lingua di insegnamento**

Italiano

---