



---

# INTRODUZIONE AL DATA MINING

INF/01 - 9 CFU - 1° semestre

## Docente titolare dell'insegnamento

### ALFREDO FERRO

**Email:** ferro@dmi.unict.it

**Edificio / Indirizzo:** Stanza 40, Blocco III, Dipartimento di Matematica e Informatica, Viale Andrea Doria 6, 95125 Catania (CT)

**Telefono:** 0957383071

**Orario ricevimento:** Su appuntamento

---

## OBIETTIVI FORMATIVI

Obiettivi formativi generali dell'insegnamento in termini di risultati di apprendimento attesi.

1. **Conoscenza e capacità di comprensione (knowledge and understanding):** Il corso mira a formare le conoscenze e le competenze di base per l'analisi, la rappresentazione, e l'organizzazione di dati.
2. **Capacità di applicare conoscenza e comprensione (applying knowledge and understanding):** lo studente acquisirà conoscenze riguardo ai modelli e gli algoritmi per l'analisi dei dati quali: mining ad alto supporto, sistemi di raccomandazione, ricerca di similarità, classificazione, clustering, text mining, network analysis.
3. **Autonomia di giudizio (making judgements):** Attraverso esempi concreti e casi di studio, lo studente sarà in grado di elaborare autonomamente soluzioni a determinati problemi legati all'analisi dei dati.
4. **Abilità comunicative (communication skills):** lo studente acquisirà le necessarie abilità comunicative e di appropriatezza espressiva nell'impiego del linguaggio tecnico nell'ambito generale dell'analisi dei dati.
5. **Capacità di apprendimento (learning skills):** il corso si propone, come obiettivo, di fornire allo studente le necessarie metodologie teoriche e pratiche per poter affrontare e risolvere autonomamente nuove problematiche che dovessero sorgere durante una attività lavorativa. A tale scopo diversi argomenti saranno trattati a lezione coinvolgendo lo studente nella ricerca di possibili soluzioni a problemi reali, utilizzando benchmark disponibili in letteratura.

## PREREQUISITI RICHIESTI

Programmazione, strutture dati, algoritmi su grafi.

---

## FREQUENZA LEZIONI

Le risorse principali messe a disposizione dello studente sono le **lezioni frontali**, la cui frequenza è **fortemente consigliata**.

Per seguire meglio le lezioni, vengono messe a disposizione le **slide** utilizzate per il corso. Le slide non costituiscono un mezzo di studio: forniscono un dettaglio puntuale sugli argomenti trattati a lezione.

---

## CONTENUTI DEL CORSO

- Background
  - Cenni su probabilità e statistica
  - Teoria spettrale
  - Entropia
- Introduzione ad R
- Data Mining ad alto supporto (apriori, insiemi frequenti)
- Data Mining a basso supporto
- Recommendation Systems
- Clustering (gerarchico, k-means, density-based)
- Classificazione (alberi decisionali, SVM, Estrattori di Regole)
- Classificatori Bayesiani
- Probabilistic Graphical Models (Catene di Markov, HMM)
- Web Mining (PageRank, Hits, Books and Authors)
- Networks (Misure di centralità, Coefficiente di Clustering)

---

## TESTI DI RIFERIMENTO

- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman, <http://www.mmds.org>
- Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber, The Morgan Kaufmann Series in Data Management Systems
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer

## ALTRO MATERIALE DIDATTICO

Il materiale didattico sarà pubblicato su [www.studium.unict.it](http://www.studium.unict.it)

---

## PROGRAMMAZIONE DEL CORSO

	* Argomenti	Riferimenti testi
1	* Introduzione al data mining: problemi, strumenti.	materiale didattico fornito dal docente
2	* Cenni di probabilità e statistica	materiale didattico fornito dal docente

---

3	* Teoria spettrale e entropia	materiale didattico fornito dal docente
4	* Introduzione ad R	materiale didattico fornito dal docente
5	* Data mining ad alto supporto	materiale didattico fornito dal docente
6	* Data mining a basso supporto	materiale didattico fornito dal docente
7	* Recommendation Systems: definizione, algoritmi, e strumenti di valutazione	materiale didattico fornito dal docente
8	* Clustering	materiale didattico fornito dal docente
9	* Classificazione: alberi decisionali e SVM	materiale didattico fornito dal docente
10	* Classificazione: estrattori di regole e classificatori bayesiani	materiale didattico fornito dal docente
11	* Predizione: regressione, regressione logistica	materiale didattico fornito dal docente
12	* Esercitazione pratica in R	materiale didattico fornito dal docente
13	* Probabilistic Graphical Models	materiale didattico fornito dal docente
14	* Web Mining	materiale didattico fornito dal docente
15	* Networks	materiale didattico fornito dal docente
16	* Cenni sul text mining	materiale didattico fornito dal docente

\* Conoscenze minime irrinunciabili per il superamento dell'esame.

**N.B.** La conoscenza degli argomenti contrassegnati con l'asterisco è condizione necessaria ma non sufficiente per il superamento dell'esame. Rispondere in maniera sufficiente o anche più che sufficiente alle domande su tali argomenti non assicura, pertanto, il superamento dell'esame.

## VERIFICA DELL'APPRENDIMENTO

### MODALITÀ DI VERIFICA DELL'APPRENDIMENTO

L'esame finale consiste in **una prova scritta**, ed un **colloquio orale** nel quale viene discusso un progetto.

La prova scritta è costituita da esercizi e domande di teoria.

Chi non supera la prova scritta, non può sostenere l'orale. La prova scritta può essere visionata prima delle prove orali.

Il progetto dovrà essere completato entro **60 giorni** dal superamento della prova scritta.

Salvo diversa comunicazione:

- l'esame scritto si svolge alle **ore 9:00**

**Note:**

- È **vietato** l'uso di qualsiasi strumento hardware (calcolatrici, tablet, smartphone, cellulari, auricolari BT etc.), di libri o documenti personali durante gli esami (scritti).
- Per sostenere gli esami è **obbligatorio prenotarsi** utilizzando l'apposito modulo del portale CEA.
- Non sono ammesse prenotazioni tardive tramite email. In mancanza di prenotazione, l'esame non può essere verbalizzato.

**PROVE IN ITINERE**

Non previste.

**PROVE DI FINE CORSO**

Un seminario di approfondimento su articoli di ricerca nel settore del data mining dovrà essere presentato dallo studente alla fine del corso, ad integrazione della votazione finale.

**ESEMPI DI DOMANDE E/O ESERCIZI FREQUENTI**

Esempi saranno pubblicati sul portale [www.studium.unict.it](http://www.studium.unict.it)

---