Versione: 16 settembre 2025, 21:37

**Università degli Studi di Salerno**
**Laurea Magistrale in Scienze Statistiche per la Finanza**
**(curriculum Big Data Analytics)**

# ADVANCED STATISTICAL LEARNING I

**Prof. Pietro Coretto**
pcoretto@unisa.it

## SYLLABUS

**Sommario.** Il corso si concentra su due temi principali: l'ensemble learning e l'analisi dei network data. Nella prima parte, si studiano le tecniche di classificazione e regressione come i Regression Trees e i metodi ensemble (bagging, random forest, boosting), con un focus sulle loro applicazioni in ambito economico e finanziario. Vengono affrontati problemi di stima, complessità dei modelli e la gestione degli iperparametri. Successivamente, il programma si sposta sull'analisi dei dati in forma di rete (network data). Vengono introdotte le basi della teoria dei grafi, le metriche di network e la costruzione di reti di correlazione. Si esaminano i metodi di community detection, come il clustering spettrale, per l'individuazione di strutture e la comprensione di fenomeni complessi, come il rischio sistemico e il contagio finanziario. L'esame finale prevede un project work e una discussione orale. Gli algoritmi e i metodi oggetto del corso saranno implementati ed illustrati mediante l'uso dell'ambiente di calcolo R.

* * *

## Indice

# 1 Organizzazione della didattica

**Lezioni.** Le lezioni si terranno nel periodo 18 settembre – 15 dicembre 2023 per un totale di 30h corrispondenti a 5cfu. Di seguito si riporta la programmazione settimanale.

```
LUN   10:10-11:50  Laboratorio SP5   [Edificio D2]
GIO   08:30-10:10  Laboratorio SP5   [Edificio D2]
```

**Avvisi e ricevimento.** Per quanto riguarda gli avvisi ed il ricevimento studenti consultare l'homepage del docente. Il calendario esami sarà disponibile attraverso la bacheca appelli di ateneo.

# 2 Materiali didattici

Slides, software, scripts, note, esercitazioni e prove d'esame saranno distribuiti durante tutta la durata del corso. La distribuzione del materiale si basa sulla piattaforma di ateneo Google Workspace. L'accesso alle risorse condivise richiede un account *@studenti.unisa.it*.

I testi indicati in bibliografia insieme al materiale fornito durante le lezioni coprono l'intero programma del corso. La valutazione del profitto riguarderà solo gli argomenti affrontati al corso, di conseguenza i testi consigliati vanno studiati solo per le parti trattate al corso.

Le lezioni si basano sull'uso del linguaggio di programmazione `R`[1] sviluppato e distributo con licenza GPL attraverso il *Comprehensive R Archive Network* (CRAN)[2]. Per questo corso è necessario installare `R-base-package` (comunemente detto `R`), ed un buon IDE.

**Installazione di `R-base-package` (seguire attentamente le istruzioni)**

- **linux:** https://cran.r-project.org/bin/linux/
- **Microsoft Windows:** https://cran.r-project.org/bin/windows/base/
- **macOS:** https://cran.r-project.org/bin/macosx/

**Installazione dell'IDE**

Vi sono tantissime alternative. Durante il corso, e nei laboratori del campus destinati alle attività didattiche, utilizzeremo `RStudio`:

- https://rstudio.com/products/rstudio/download/#download

Hastie et al. (2009) è il libro di testo principale del corso, sebbene il testo di Efron e Hastie (2016) sarà molto utile per avere una prima visione di insieme dei singoli argomenti. Tuttavia, sarà necessario integrare la bibliografia consigliata con i materiali forniti durante le lezioni. La valutazione del profitto riguarderà solo gli argomenti affrontati al corso, di conseguenza i testi consigliati vanno studiati solo per le parti trattate al corso.

[1] Vedi: https://it.wikipedia.org/wiki/R_(software)
[2] Vedi: https://en.wikipedia.org/wiki/R_package#Comprehensive_R_Archive_Network_(CRAN)

# 3 Programma dettagliato

Di seguito si riporta il dettaglio dei contenuti del corso. La trattazione degli argomenti si basa sempre su applicazioni nell'ambito della finanza. La maggior parte delle lezioni includeranno una componente metodologica ed una componente computazionale ricca di esempi e casi studio.

**Previsione ed ottimalità delle previsioni.** Ruolo dei metodi di learning nelle applicazioni finanziare – breve panoramica sui temi del corso – supervised vs unsupervised learning – richiami di teoria della probabilità: bernoullizzazione di un esperimento e valori attesi condizionati – il problema della classificazione – nozione di rischio – loss 0–1 – loss quadratica – teorema del previsore ottimale e sue applicazioni alla teoria della previsione – previsioni ottimali nel caso di loss 0–1 e loss quadratica. – il problema della gestione della complessità ottimale nel contesto dei modelli per la previsione.

**Ensemble learning e classificazione.** Classificazione e principali applicazioni alla finanza – regressione nonparametrica – alberi di regressione (RT, regression-trees) – regressione e classificazione con RT – principali applicazioni applicazioni economiche e finanziarie – il problema di stima dei RT – recursive binary splitting – il problema della complessità nei modelli di regressione su alberi – adaptive pruning – metodi di resampling: cross-validation e bootstrap – kFCV e out-of-bag bootstrap per la scelta della complessità ottimale del previsore – problemi dei RT: discontinuità, instabilità e varianza di previsione – bootstrap resampling – bagging: algoritmo e sue varianti – consensus bagging e bagged posteriors – data splitting e correlazione dei previsori – random forest – gestione degli iperparametri nel random forest – CART con dati pesati e boosting – algoritmo AdaBoost – scelta della lunghezza del tree e del *base learaner* – generalizzazioni del boosting – feature importance – vantaggi e limiti dei metodi ensemble.

**Network data e community detection.** I networks come rappresentazioni di strutture dati complesse – principali applicazioni alla finanza – introduzione ai grafi e network complessi: nodi, archi, grafi diretti e non diretti, pesati – matrici di rappresentazione: matrice di adiacenza e matrice Laplaciana – metriche di network: grado di un nodo, centralità (betweenness, closeness, eigenvector) – fonti dati: dati di prezzo e rendimento nelle loro varie forme – costruzione di network di correlazione e di correlazione parziale – il problema della stima di grandi matrici di correlazione – visualizzazione di network – community detection e partizionamento di un grafo – problemi Mincut, MaxFlow e rilassamento – introduzione al Clustering Spettrale su grafi – utilizzo dei metodi network per lo studio del rischio di contagio finanziario e del rischio sistemico, ottimizzazione di portafoglio e l'individuazione dei segmenti.

# 4 Esame e valutazione del profitto

L'esame prevede l'elaborazione autonoma di un project work e una prova orale.

**Project Work.** Lo studente dovrà consegnare il materiale (solitamente un manoscritto e software) al docente via email almeno 3 giorni prima della data d'esame. Il project work dovrà avere requisiti specifici indicati dal docente durante il corso. Il project work viene valutato con un massimo di 15 punti.

**Discussione Orale.** L'esame si completa con una prova orale della durata approssimativa di 45 minuti. Partendo dalla discussione del tema del project work, durante la prova orale sarà verificata la preparazione su tutti i contenuti del corso. La prova sarà valutata con un massimo di 15 punti.

Il voto finale, espresso in trentesimi, consiste nella somma delle due valutazioni. La lode viene attribuita quando lo studente dimostra una padronanza tale da saper estendere le conoscenze acquisite alla risoluzione autonoma di problemi non trattati a lezione.

# 5 Riferimenti bibliografici

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.