UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Science: Longitudinal, Multilevel and Multivariate Analysis

2627-1-F8804N008

Obiettivi

Conoscenza e comprensione: acquisire una conoscenza approfondita delle tecniche avanzate di data science applicabili all'analisi dei fenomeni sociologici.

Capacità di applicare conoscenza e comprensione: fornire le competenze necessarie per analizzare basi di micro dati cross-sectional ripetuti, longitudinali panel e muldidimensionali in ambito sociologico per rispondere ad interrogativi di natura associativa, predittiva e causale.

Autonomia di giudizio: sviluppare la capacità di valutare in modo critico i metodi e i risultati delle analisi quantitative nella ricerca sociologica, riconoscendone i presupposti teorici, i limiti e le implicazioni pratiche.

Abilità comunicative: promuovere la capacità di strutturare e argomentare efficacemente i risultati quantitativi all'interno di una cornice teorica sociologica.

Capacità di apprendere: promuovere l'autonomia nello studio e nella ricerca, stimolando la capacità di approfondire in modo critico e indipendente i contenuti del corso.

Contenuti sintetici

Il corso fornisce una panoramica avanzata degli strumenti per l'analisi di dati complessi in ambito sociologico, con particolare attenzione a dati gerarchici, longitudinali (panel) e multivariati. Tra gli argomenti trattati figurano: le tecniche di regressione multilivello, i modelli econometrici per dati panel, le equazioni strutturali (SEM), i modelli di event history analysis, le principali tecniche di riduzione della dimensionalità (cluster analysis e analisi fattoriale), nonché alcune tecniche di machine learning supervisionato e non supervisionato, come le reti neurali.

Programma esteso

La prima parte del corso è dedicata ai metodi per l'analisi di dati gerarchici caratterizzati da strutture di varianza complesse. In particolare, verranno introdotti i modelli multilivello, che rappresentano un'estensione dei modelli di regressione tradizionali per dati organizzati gerarchicamente. Queste tecniche risultano applicabili anche ai dati longitudinali di tipo panel, nei quali un medesimo outcome è osservato in più occasioni temporali all'interno delle stesse unità analitiche. La prima parte si conclude con l'esposizione delle principali tecniche econometriche per l'analisi di dati panel: il modello a effetti fissi, il modello a effetti casuali e lo stimatore Differences-in-Differences (DiD). La seconda parte del corso si concentrerà sull'analisi di dati multivariati. Verranno presentate le principali tecniche di riduzione della dimensionalità, tra cui l'analisi delle componenti principali (PCA), l'analisi fattoriale e l'analisi dei gruppi (cluster analysis di tipo gerarchico e non gerarchico). A seguire, verranno introdotti i modelli di equazioni strutturali (SEM), che integrano in un unico approccio la logica della regressione causale e quella dell'analisi fattoriale. Saranno inoltre affrontati i modelli di Event History Analysis, con particolare riferimento allo stimatore di Kaplan-Meier e al modello di regressione di Cox. Il corso si chiuderà con un approfondimento dedicato a specifiche tecniche di machine learning, in particolare alle reti neurali supervisionate (Multilayer Perceptron) e non supervisionate (Self-Organizing Map)..

Prerequisiti

Prerequisiti per il corso includono la conoscenza dei modelli di regressione lineare e logistica e una preparazione teorico-metodologica di base nell'ambito della ricerca sociale.

Modalità didattica

Il corso prevede un totale di 56 ore di didattica in presenza, articolate in lezioni che combinano modalità espositiva e attività interattive. Ciascun incontro si compone di una prima parte dedicata alla presentazione dei contenuti teorici e metodologici (modalità frontale) e di una seconda parte orientata alla partecipazione attiva degli studenti e delle studentesse attraverso esercitazioni individuali o di gruppo, presentazioni e momenti di discussione collettiva. Complessivamente, circa il 70% delle ore sarà destinato alla didattica frontale, mentre il restante 30% sarà dedicato ad attività laboratoriali e interattive. Il corso si svolge in lingua italiana e le esercitazioni verranno svolte utilizzando il software statistico Stata.

Materiale didattico

Diapositive e materiali didattici a cura del docente.

Kreft, I. G. G., & de Leeuw, J. (1998). Introducing multilevel modeling. Thousand Oaks, CA: Sage Publications. (opzionale)

Longhi, S., & Nandi, A. (2015). A practical guide to using panel data. London: SAGE Publications Ltd. (opzionale)

Singer, J.D. & Willett, J.B. (2003), *Applied Longitudinal Data Analysis* (*ALDA*), Oxford University Press. (opzionale)

De Lillo, A., Argentin, G., Lucchini, M., Sarti, S., & Terraneo, M. (2007). L'analisi multivariata per le scienze sociali (Cap. 7–8). Milano: Pearson Education. (opzionale)

Periodo di erogazione dell'insegnamento

febbraio 2027 - maggio 2027

Modalità di verifica del profitto e valutazione

Lo studente potrà optare per una prova orale, basata sui materiali messi a disposizione dal docente e indicati in bibliografia. Durante l’esame verranno poste sei domande, ciascuna riferita agli argomenti trattati nel corso. Ogni risposta sarà valutata con un punteggio compreso tra 0 e 6. Il voto complessivo sarà determinato dalla somma dei punteggi ottenuti per ciascuna risposta.
Qualora la somma risulti inferiore a 18, la prova sarà considerata insufficiente. In caso di punteggio pari o superiore a 31, il voto finale sarà 30 con lode.
La valutazione delle risposte si baserà su tre criteri fondamentali: correttezza, completezza e chiarezza espositiva.
In alternativa, lo studente può sostenere una prova scritta in aula, utilizzando il proprio computer personale e il software Stata. La prova consisterà nell’implementazione di quattro modelli di analisi dei dati tra quelli presentati durante il corso. Ciascuna risposta sarà valutata con un punteggio compreso tra 0 e 8. Il voto complessivo sarà determinato dalla somma dei punteggi ottenuti per ciascun quesito. Se il punteggio totale sarà inferiore a 18, la prova sarà considerata insufficiente; se pari o superiore a 31, il voto finale sarà 30 e lode.
La valutazione si baserà su tre criteri principali: correttezza dell’analisi, completezza della risposta e chiarezza espositiva.
Durante la prova, il docente fornirà i dataset necessari per lo svolgimento delle analisi. Il tempo a disposizione per completare la prova scritta sarà di 120 minuti.

Orario di ricevimento

Mercoledi' 11.00-12.00

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ