UNIVERSITÀ DI PISA

STATISTICS FOR DATA SCIENCE

SALVATORE RUGGIERI

Anno accademico	2023/24
CdS	DATA SCIENCE AND BUSINESS INFORMATICS
Codice	628PP
CFU	9

Moduli	Settore/i	Tipo	Ore	Docente/i
STATISTICS FOR DATA SCIENCE	SECS-S/01	LEZIONI	72	SALVATORE RUGGIERI

Obiettivi di apprendimento

Conoscenze
Lo studente che completerà con successo il corso avrà una solida conoscenza dei principali metodi e strumenti software per l'analisi statistica, con particolare riferimento ai concetti di modello statistico, di inferenza dei parametri e di intervalli di confidenza, di test di ipotesi, di inferenza Bayesiana e causale, con applicazione specifica ai problemi e contesti utili nella data science. Lo studente sarà in grado di utilizzare il linguaggio di programmazione R per condurre analisi statistiche dei dati.

Modalità di verifica delle conoscenze
Gli studenti saranno valutati sulla comprensione dei contenuti del corso, sull'uso della terminologia, sulla formalità del ragionamento, in particolare nell'applicare i metodi statistici in contesti applicativi.

Capacità
Gli studenti saranno in grado di applicare i concetti di analisi statistica e le metodologie più appropriate a casi di studio tipici della data science. Gli studenti saranno in grado di utilizzare il linguaggio R per analisi statistiche.

Modalità di verifica delle capacità
Gli studenti frequentanti potranno svolgere un progetto in gruppo di natura analitica (su un dataset di grandi dimensioni) o di ricerca (comprensione e riproduzione di risultati scientifici), rispondendo ad alcuni obiettivi di analisi. Il progetto permetterà di verificare la capacità di analisi dei problemi e di (re-)implementazione di soluzioni basate su modelli statistici.

Comportamenti
Gli studenti saranno in grado di valutare bias (distorsioni) dei modelli statistici, in particolare in contesti applicazioni con possibilità di ingiusta discriminazione di persone o gruppi sociali.

Modalità di verifica dei comportamenti
Il progetto in gruppi e la verifica orale discuteranno eventuali aspetti di bias e discriminazione nell'uso dei modelli statistici.

Prerequisiti (conoscenze iniziali)
Conoscenza di base di analisi matematica e ottimizzazione. Una conoscenza di base di calcolo delle probabilità è utile, ma non strettamente indispensabile.

Indicazioni metodologiche
Modalità di erogazione: in presenza
Attività di apprendimento:

- lezioni frontali e esercitazioni in classe
- partecipazione alle discussioni
- studio individuale

- progetto di gruppo

Frequenza: fortemente consigliata, ma non obbligatoria
Metodologie di insegnamento:

- lezioni frontali
- esercitazioni nel linguaggio R

Programma (contenuti dell'insegnamento)

Il programma ricomprende le metodologie, le tecniche e gli strumenti base dell'analisi statistica. Sono ricompresi: calcolo delle probabilità, variabili aleatorie, teoremi di convergenza, modelli statistici, teoria della stima, test di ipotesi, inferenza Bayesiana, ragionamento causale. Ulteriori argomenti inclusi sono il metodo bootstrap, l'expectation-maximization, e le applicazioni ai problemi di data science. Il programma copre anche l'uso del linguaggio di programmazione R per l'analisi statistica.

Bibliografia e materiale didattico

- F.M. Dekking C. Kraaikamp, H.P. Lopuha, L.E. Meester. **A Modern Introduction to Probability and Statistics**. Springer, 2005.
- P. Dalgaard. **Introductory Statistics with R**. 2nd edition, Springer, 2008.
- Capitoli o articoli scientifici selezionati dal docente.

Indicazioni per non frequentanti

Gli studenti non frequentanti non potranno svolgere il progetto. Tutte le altre modalità sono invariate.

Modalità d'esame

L'esame consiste di una prova scritta e di una prova orale. La prova scritta dura 2 ore ed include domande a risposta aperta ed esercizi (sia sulla teoria che sul linguaggio R). Ciascuna domanda o esercizio contribuisce con un punteggio al totale di 30 punti. Esempi di testi scritti e soluzioni sono pubblicate sulla pagina web del corso. Gli studenti che ottengono almeno 18/30 sono ammessi alla prova orale. La prova orale consiste di domande aperte sugli argomenti del corso e sull'uso del linguaggio R. Gli studenti che frequentano regolarmente le lezioni potranno sostituire la prova scritta con un progetto di gruppo da svolgere durante il semestre.

Pagina web del corso

http://didawiki.di.unipi.it/doku.php/mds/sds/start

Ultimo aggiornamento 26/07/2023 10:10