UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Scienze Economiche, Aziendali e Statistiche		
ANNO ACCADEMICO OFFERTA	2024/2025		
ANNO ACCADEMICO EROGAZIONE	2024/2025		
CORSO DILAUREA MAGISTRALE	STATISTICA E DATA SCIENCE		
INSEGNAMENTO	MODELLI STATISTICI E DATA MINING C.I.		
CODICE INSEGNAMENTO	23831		
MODULI	Si		
NUMERO DI MODULI	2		
SETTORI SCIENTIFICO-DISCIPLINARI	SECS-S/01		
DOCENTE RESPONSABILE	CHIODI MARCELLO	Professore Ordinario	Univ. di PALERMO
ALTRI DOCENTI	PLAIA ANTONELLA CHIODI MARCELLO	Professore Ordinario Professore Ordinario	Univ. di PALERMO Univ. di PALERMO
CFU	12		
PROPEDEUTICITA'			
MUTUAZIONI			
ANNO DI CORSO	1		
PERIODO DELLE LEZIONI	1° semestre		
MODALITA' DI FREQUENZA	Facoltativa		
TIPO DI VALUTAZIONE	Voto in trentesimi		
ORARIO DI RICEVIMENTO DEGLI STUDENTI	**CHIODI MARCELLO** Martedì 15:00 17:00 stanza del docente (edificio 13); eccezionalmente su teams Venerdì 12:00 13:00 stanza del docente (edificio 13); eccezionalmente su teams **PLAIA ANTONELLA** Mercoledì 15:00 17:00 La modalita, in studio o su Teams, va concordata col docente		

DOCENTE: Prof. MARCELLO CHIODI

PREREQUISITI	Conoscenza dei fondamenti e metodi dell'inferenza statistica classica (al livello del Corso di Inferenza Statistica STAD) e dell'inferenza sui Modelli Lineari (al livello del Corso di Modelli Lineari STAD); conoscenza dell'ambiente di programmazione statistica R a livello intermedio o altri software scientifici opensource (p.e. Python).
RISULTATI DI APPRENDIMENTO ATTESI	1. Conoscenza e capacita' di comprensione 1.1. Conoscenza dei metodi avanzati dell'inferenza statistica classica (basata sull'approccio di verosimiglianza). 1.2. Conoscenza dei metodi di base dell'inferenza Bayesiana. 1.3. Comprensione delle giustificazioni teoriche nel caso di modelli a più parametri, con approfondimenti teorici e con tecniche di simulazione 2. Capacita' di applicare conoscenza e comprensione 2.1. Capacita' di specificare il modello statistico con un approccio critico, partendo dagli obbiettivi conoscitivi/operativi dello studio esaminato. 2.2. Capacita' di usare in modo integrato le conoscenze acquisite in corsi precedenti per trattare problemi applicativi reali, inclusi problemi non-standard. 2.3. Capacita' di dimostrare risultati teorici in modo formale. 3. Autonomia di giudizio 3.1. Comprensione critica delle caratteristiche, potenzialita' e limiti di modelli statistici gia' conosciuti, e capacita' di arricchirli con estensioni e nuove caratteristiche quando necessario. 4. Abilita' comunicative 4.1. Capacita' di discutere le caratteristiche di un dato problema inferenziale, sia con statistici che con non-statistici. 4.2. Capacita' di scrivere un rapporto tecnico-scientifico, focalizzato sul modello statistico scelto e sull'interpretazione sostantiva dei risultati. 5. Capacita' d'apprendimento 5.1. Capacita' di utilizzare le nozioni e competenze acquisite in successivi corsi di Statistica e Statistica Applicata e nella tesi finale. 5.2. Capacita' di consultare e comprendere la letteratura statistica internazionale, allo scopo di aggiornare le proprie conoscenze teoriche e competenze tecniche.
VALUTAZIONE DELL'APPRENDIMENTO	Prova finale scritta e orale. La prova scritta consiste nell'analisi di un dataset reale in laboratorio informatico, con utilizzo dell'ambiente di programmazione statistica R. Il candidato ha di norma tre ore a disposizione, alla fine delle quali deve consegnare un rapporto tecnico finale. La prova scritta ha come esito solo due possibili risultati: "Ammesso alla prova orale" vs. "Non ammesso alla prova orale". La condizione necessaria per il superamento della prova scritta e' che il candidato dimostri una sufficiente capacita' di: (i) utilizzare in modo autonomo e critico i metodi statistici appresi a lezione per analizzare gli specifici problemi che caratterizzano il dataset proposto; (ii) interpretare i risultati statistici raggiunti; (iii) scrivere in modo efficace un rapporto tecnico-scientifico. La prova orale, cui sono ammessi solo gli studenti che abbiano superato la prova scritta, si articola in due fasi: (i) la discussione del rapporto tecnico finale redatto dal candidato nella prova scritta; (ii) la verifica della conoscenza e capacita' del candidato di illustrare e discutere i principali risultati teorici presentati nelle lezioni frontali. In caso di superamento, il voto finale (espresso nel campo di variazione 18/30 - 30/30, piu' l'eventuale lode) riflettera': (i) il livello mostrato dal candidato, nella prova scritta di laboratorio, di raggiungimento dei "Risultati di apprendimento attesi", con particolare riferimento alle voci sub. 2 e 4.2 fino ad un massimo di 15/30; (ii) il livello mostrato dal candidato, nella prova orale, di raggiungimento dei "Risultati di apprendimento attesi", con particolare riferimento alle voci 1.1, 1.2, 1.3, 2.3, 4.1 (fino ad un massimo di 15/30). Il voto finale sara' ottenuto per somma delle due componenti ora descritte. Per superare l'esame, e ottenere quindi un voto non inferiore a 18/30, lo studente deve dimostrare un livello sufficiente di raggiungimento dei "Risultati di apprendimento attesi" sia nella prova scritta che in quella orale. Per conseguire la valutazione di 30/30, lo studente deve dimostrare un livello ottimo di raggiungimento dei "Risultati di apprendimento attesi" sia nella prova scritta che in quella orale. La lode e' riservata agli studenti che dimostrano una padronanza eccellente dei contenuti del corso ed uno spiccato senso critico nel loro utilizzo.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali, esercitazioni in laboratorio informatico, analisi di casi di studio reali.

MODULO
MODELLI STATISTICI

Prof. MARCELLO CHIODI

TESTI CONSIGLIATI

a)appunti di lezione (lecture notes);
b)Agresti, A., (2015) Foundations of Linear and Generalized Linear Models- Wiley eds.
c) Mc Cullagh, Nelder, (1989) Generalized Linear Models-Chapman and Hall eds.
d) Wood, S. (2006) , Generalized Additive Models_ An Introduction with R- Chapman and Hall
e) Pawitan, Y. (2001) In All Likelihood. Oxford Science Publications, Oxford

TIPO DI ATTIVITA'	B
AMBITO	84542-Discipline Statistiche
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	108
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	42

OBIETTIVI FORMATIVI DEL MODULO

Questo corso mira ad arricchire il bagaglio teorico ed applicativo dello studente nella costruzione e interpretazione dei modelli statistici, approfondendo le unita' didattiche: (a) gli sviluppi in ambito di modelli di tipo regressivo non lineare (GLM ed estensioni); (b) Approfondimento di alcuni aspetti dell'inferenza

parametrica classica; (c) cenni all'inferenza Bayesiana; (d) Cenni alla selezione del modello con riferimento alle capacità descrittive e/o predittive. La parte teorica, erogata nelle lezioni frontali, sara' integrata dal punto di vista applicativo nelle esercitazioni in laboratorio, realizzate nell'ambiente statistico R. Dopo aver frequentato questo corso con successo, gli studenti preparati dovrebbero essere capaci di:
(i) specificare un modello statistico appropriato per i dati in esame (GLM o altri modelli), fare inferenza su tale modello e interpretare i risultati;
(ii) riconoscere situazioni in cui e' necessario ricorrere ad una estensione dei GLM standard, e fare inferenza su tali modelli estesi;
(iii) avere un approccio critico al processo di costruzione dei modelli; (iv) sviluppare competenze di base sull' inferenza Bayesiana

PROGRAMMA

ORE	Lezioni
8	(a) Richiami sui modelli lineari, ordinari e generali, predittori lineari e configurazione della matrice del disegno. La distribuzione normale multivariata. Teoria asintotica dell'inferenza su più parametri nel caso regolare
4	Approcci generali all'inferenza: Cenni all'inferenza Bayesiana. Distribuzioni a priori e a posteriori; il ruolo della verosimiglianza. Stima Bayesiana puntuale e intervallare.
12	I modelli lineari generalizzati. Ruoli diversi dei vari elementi: predittore lineare, funzione legame, distribuzione appartenente alla famiglia esponenziale. Metodi numerici per la stima dei parametri (IWLS). Proprietà asintotiche. Residui, strumenti di diagnostica. Confronto fra modelli. Selezione di modelli. Aspetti computazionali ORE

ORE	Esercitazioni
14	Sviluppi in tema di modelli di tipo regressivo: esercitazioni di laboratorio con R. GLM: esempi su varie distribuzioni, casi di studio, software R e package vari. Stima dei parametri, interpretazione dei risultati, confronto fra modelli, diagnostica. Qualche applicazione di tecniche di simulazione
4	Sviluppi in tema di inferenza: esercitazioni di laboratorio con R.

MODULO
DATA MINING

Prof.ssa ANTONELLA PLAIA

TESTI CONSIGLIATI

Dispense rese disponibili dal docente sul portale di Ateneo. Risorse on-line indicate dal docente durante il corso.
Breiman, L. Friedman, J. H. Olshen, R. A. Stone, C. J. (1984) Classification and regression trees, Chapman & Hall. Capp. 1-5, 8
G. James, D. Witten, T. Hastie, R. Tibshirani . (2013) An Introduction to Statistical Learning, with applications in R. Springer. Cap. 8
Stef van Buuren, (2012) Flexible Imputation of Missing Data, Chapman & Hall, capp 1-4, 7.2

TIPO DI ATTIVITA'	B
AMBITO	84542-Discipline Statistiche
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	108
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	42

OBIETTIVI FORMATIVI DEL MODULO

Il corso illustra metodi statistici di apprendimento da dati empirici complessi.
L'obiettivo principale e' l'analisi di grandi database al fine di trovare pattern, associazioni, cambiamenti, anomalie e strutture di particolare interesse.
Alla fine del corso il discente sara' in grado di individuare gli strumenti adeguati per l'analisi che deve svolgere e applicarli, sintetizzando e riportando in report e presentazione i risultati in modo efficace.
Versione inglese

PROGRAMMA

ORE	Lezioni
4	Analisi esplorativa dei dati
4	Integrazione di dati da fonti diverse
4	Rappresentazioni grafiche di dati multidimensionali
2	Web scraping
4	Trattamento dei dati mancanti
6	Alberi decisionali e ensemble methods

ORE	Esercitazioni
4	Integrazione di dati da fonti diverse
4	Rappresentazioni grafiche di dati multidimensionali
2	Web scraping
4	Trattamento dei dati mancanti
4	Alberi decisionali e ensemble methods