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	DATA AND TEXT MINING C.I.		
CODICE INSEGNAMENTO	23844		
MODULI	Si		
NUMERO DI MODULI	2		
SETTORI SCIENTIFICO-DISCIPLINARI	SECS-S/01		
DOCENTE RESPONSABILE	PLAIA ANTONELLA	Professore Ordinario	Univ. di PALERMO
ALTRI DOCENTI	PLAIA ANTONELLA	Professore Ordinario	Univ. di PALERMO
CFU	9		
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	**PLAIA ANTONELLA** Mercoledì 15:00 17:00 La modalita, in studio o su Teams, va concordata col docente		

DOCENTE: Prof.ssa ANTONELLA PLAIA

PREREQUISITI	Fondamenti di statistica descrittiva e inferenziale. Elementi di calcolo delle probabilità . Conoscenza del software R
RISULTATI DI APPRENDIMENTO ATTESI	Conoscenza e capacita' di comprensione Conoscenza dei metodi e delle procedure statistiche per analisi esplorative di data-set di grandi dimensioni. Capacita' di applicare conoscenza e comprensione Lo studente dovra' essere in grado di: 1. formulare correttamente un problema 2. scegliere soluzioni di analisi statistico-informatiche appropriate, 3. tradurre i risultati in decisioni operative. Autonomia di giudizio Lo studente dovra' essere in grado di: 1. tradurre in termini statistici una esigenza conoscitiva sorta in differenti campi applicativi 2. intervenire con attivita' di pulizia, riorganizzazione, analisi descrittiva e interpretazione, dei dati, 3. elaborare e comunicare coerentemente i risultati delle proprie analisi. Abilita' comunicative Lo studente dovra' essere in grado di: 1. comunicare con chiarezza, sia agli specialisti che ai non specialisti della materia, i concetti, e le tecniche di analisi dei dati studiati, 2. giustificare le scelte degli strumenti utilizzati per l'analisi, 3. comunicare i risultati con linguaggio appropriato. Capacita' d'apprendimento Lo studente avra' sviluppato le abilita' di apprendimento necessarie per approfondire autonomamente lo studio delle tecniche statistiche piu' comunemente utilizzate nell'analisi di grandi moli di dati.
VALUTAZIONE DELL'APPRENDIMENTO	Presentazione di due report e prova orale. La verifica dell'apprendimento avviene attraverso: - Per chi sostiene l'esame durante la sessione successiva a quella di erogazione del corso 1. Consegna di un report intermedio contenente l'analisi di un dataset assegnato durante la sesta settimana del corso a gruppi di studenti composti da al piu' 3 persone. La consegna e' prevista al rientro della pausa infrasemestre, e la presentazione dei risultati dell'analisi, da parte di un rappresentante per ogni gruppo, all'inizio dell'ottava settimana di corso. 2. Consegna di un report finale contenente l'analisi di un dataset assegnato alla fine del corso (consegna almeno una settimana prima della prova orale) e sua presentazione (in power point o altro). - Per chi sostiene l'esame in altre sessioni Contattare il docente almeno 1 mese prima dell'inizio della sessione di esami (primaverile, estiva, autunnale) per avere assegnata la prova. 1. Consegna di un report intermedio contenente l'analisi di un dataset assegnato almeno 3 settimane prima dell'inizio della sessione. La consegna e' prevista 1 settimana prima del primo appello della sessione, e la presentazione dei risultati dell'analisi durante uno degli appelli. 2. Consegna di un report finale contenente l'analisi di un dataset assegnato almeno 3 settimane prima dell'inizio della sessione (consegna almeno una settimana prima della prova orale) e sua presentazione (in power point o altro). Sia il report finale che quello intermedio (al piu' 4 pagine, piu' eventuali grafici e tabelle) devono essere consegnati, accompagnati dai corrispondenti codici in R. La prova orale consistera' nella presentazione (in power point o altro) dell'analisi e dei risultati descritti nel report finale, per valutare meglio le conoscenze e abilita' possedute dallo studente, nonche' la sua capacita' di trasmetterle con idoneo linguaggio statistico. VALUTAZIONE FINALE La valutazione finale dell'esame prendera' in considerazione tre aspetti: i) la padronanza degli argomenti; ii) la capacita' di applicazione delle conoscenze e iii) la proprieta' di linguaggio, valutati nel complesso dei report intermedio (peso 0,3) e finale (peso 0,5) e della prova orale (peso 0,2). La valutazione sarà sufficiente se il candidato avrà scelto i metodi corretti di pulizia e riorganizzazione dei dati forniti per entrambi i report e individuato gli strumenti di analisi appropriati, anche se l'analisi non risulterà efficiente. Per una valutazione sufficiente, inoltre, il candidato deve dimostrare una sufficiente capacità argomentativa ed espositiva durante la prova orale. Quanto piu' l'esaminando darà evidenza, nella scrittura dei report e nella prova orale, delle sue capacita' argomentative ed espositive, nonche' di proprieta' di linguaggio statistico, e di uso efficiente del linguaggio di programmazione R per

	l'analisi dei dati, tanto piu' la valutazione sara' positiva. La Commissione giudicatrice e' formata dal docente titolare dell'insegnamento (Presidente) e da almeno un altro docente, Professore o Ricercatore, del medesimo o affine settore disciplinare, o un cultore della materia.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali, Esercitazioni in aula di informatica Durante il corso, il docente condividerà con gli studenti un breve articolo, un capitolo di libro, o una sua parte in lingua inglese di carattere divulgativo, che sarà oggetto di analisi e dibattito anche finalizzati ad una presentazione o discussione autonoma da parte degli studenti

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	70296-Formazione matematico-statistica
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

MODULO
TEXT 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.
“Text Mining with R by Julia Silge and David Robinson (O’Reilly). Copyright 2017 Julia Silge and David Robinson, 978-1-491-98165-8.” https://www.tidytextmining.com/index.html
Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

TIPO DI ATTIVITA'	B
AMBITO	70296-Formazione matematico-statistica
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	54
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	21

OBIETTIVI FORMATIVI DEL MODULO

Il corso illustra metodi statistici di apprendimento da dati empirici complessi con particolare riferimento ai dati testuali.
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.

PROGRAMMA

ORE	Lezioni
3	Fondamenti del Text Mining: il formato "tidy".
3	Relazioni tra le parole: N-grammi, correlazioni, e network testuali.
3	Introduzione alla Sentiment Analysis
3	Topic models

ORE	Esercitazioni
3	Analisi esplorative e rappresentazioni grafiche dei dati testuali
3	Applicazione delle tecniche analisi testuale e sentiment alle canzoni
3	Topic models