

Principali informazioni sull'insegnamento	
Titolo insegnamento	Modellizzazione Statistica
Corso di studio	Laurea Magistrale in Data Science
Crediti formativi	6
Denominazione inglese	Statistical Modelling
Obbligo di frequenza	No
Lingua di erogazione	Italiano

Docente responsabile	Nome Cognome	Indirizzo Mail
	Massimo Bilancia	massimo.bilancia@uniba.it massi.bilancia@gmail.com

Dettaglio credi formativi	Ambito disciplinare	SSD	Crediti
	Scienze Economiche e Statistiche	SECS-S/01 (Statistica)	6 (4 T1 + 2 T2)

Modalità di erogazione	Lezioni frontali (4 CFU di tipo T1) Lezioni di laboratorio ed esercitazioni (2 CFU di tipo T2)
Periodo di erogazione	II° semestre
Anno di corso	Primo
Modalità di erogazione	Lezioni frontali Lezioni di laboratorio

Organizzazione della didattica	
Ore totali	150
Ore di corso	62
Ore di studio individuale	88

Calendario	
Inizio attività didattiche	1 marzo 2021
Fine attività didattiche	4 giugno 2021

Syllabus	
Prerequisiti	Elementi di base della statistica descrittiva, del calcolo delle probabilità e dell'inferenza statistica
Risultati di apprendimento previsti	<ul style="list-style-type: none"> • <i>Conoscenza e capacità di comprensione</i> Conoscenza dei fondamenti statistico-probabilistici dell'apprendimento supervisionato, e dei principali modelli statistici utilizzati per l'apprendimento supervisionato, nonché dei principali modelli utilizzati a tal fine. • <i>Conoscenza e capacità di comprensione applicate</i> Introduzione ai fondamenti della programmazione in R che costituisce, insieme a Python, la piattaforma di Data Science attualmente più utilizzata. In particolare, di R lo studente

	<p>dovrà essere in grado di utilizzare la versione 'moderna' basata su RStudio + Tidyverse</p> <ul style="list-style-type: none"> • Autonomia di giudizio Capacità di individuare autonomamente il modello più adeguato all'analisi del problema, e di implementarlo mediante la piattaforma di Data Science utilizzata durante lo svolgimento del corso (R + RStudio + Tidyverse). • Abilità comunicative Lo studente dovrà essere in grado di presentare i risultati prodotti tramite R Markdown, che costituisce una delle soluzioni più utili per la presentazione e la reportistica dinamica delle analisi effettuate. • Capacità di apprendere Acquisizione di un insieme di conoscenze teoriche e pratiche che mettano in grado lo studente di leggere autonomamente la letteratura più recente nell'ambito del Machine Learning.
Contenuti di insegnamento	<ul style="list-style-type: none"> - Inferenza frequentista e Bayesiana - Modello di regressione lineare - Modello di regressione logistica - Classificazione - Regolarizzazione e regressione non-parametrica - Teoria dell'apprendimento supervisionato

Programma	<p>Teoria (32 h)</p> <p>Richiami di calcolo delle probabilità – Metodi Monte Carlo – Inferenza frequentista e stime di massima verosimiglianza: proprietà in senso frequentista e metodi numerici – Inferenza Bayesiana e modelli gerarchici – Metodi computazionali per l'inferenza Bayesiana: Gibbs sampling e Markov Chain Monte Carlo (MCMC) – Modelli di regressione di classificazione – Regressione lineare semplice e multipla: inferenza frequentista e Bayesiana – Regressione logistica semplice e multinomiale – Metodi numerici per la stima dei parametri: discesa del gradiente, IRLS e cenni sull'inferenza Bayesiana – Interpretazione dei parametri: odds ratio – Classificazione logistica – Analisi discriminante lineare, quadratica e kNN – Classificatori generativi e discriminativi – Regressione non parametrica: basi polinomiali e splines, smoothing splines – Selezione automatica del parametro di smoothing – Modelli additivi generalizzati (GAM) – Apprendimento supervisionato e non-supervisionato – Teoria delle decisioni – Il teorema di trade-off bias-varianza per i modelli di regressione – Il trade-off bias-varianza nei problemi di classificazione – Inferenza per l'errore di generalizzazione: Cp di Mallows, AIC e BIC, Cross-Validation, Bootstrap – Regolarizzazione: Ridge Regression, LASSO ed Elastic Net – La determinazione del modello ottimale in un'ottica Bayesiana – Il framework PAC e la dimensione di Vapnik-Chervonenkis – Metriche per la</p>
------------------	--

	<p>valutazione dell'accuratezza sul test set.</p> <p>Laboratorio (30 h)</p> <p>Fondamenti di programmazione in R – RStudio/Tidyverse – Le librerie di Tidyverse - La libreria caret per l'automazione dei flussi di lavoro – R Markdown – Stan e Greta per l'inferenza Bayesiana in R</p>
Testi di riferimento	<p>RIFERIMENTO BASE PER LA TEORIA</p> <p>- James G., Witten D., Hastie T., Tibshirani R. (2020) Introduzione all'apprendimento statistico. Con applicazioni in R. Piccin-Nuova Libreria. ISBN-10/ASIN: ISBN-10: 8829930946. ISBN-13: 978-8829930944. La prima edizione in inglese è liberamente scaricabile al seguente indirizzo: https://www.statlearning.com/</p> <p>TESTI DI APPROFONDIMENTO</p> <p>- K. Murphy (2012): Machine Learning. A Probabilistic Perspective. The MIT Press, 1st Edition. ISBN-10/ASIN: 0262018020. ISBN-13: 978-0262018020.</p> <p>- T. Hastie, R. Tibshirani, J. Friedman (2013). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Verlag, 2nd Edition. ISBN-10/ASIN: 0387848576. ISBN-13: 978-0387848570. Scaricabile liberamente da: https://web.stanford.edu/~hastie/Papers/ESLII.pdf</p> <p>RIFERIMENTO BASE PER IL LABORATORIO</p> <p>- G. Golemund, H. Wickam (2017). R for Data Science. O'Reilly Media, 1st Edition. ISBN-10/ASIN:1491910399. ISBN- 13: 978-1491910399 .</p>
Note ai testi di riferimento	I testi di riferimento saranno supportati da un insieme di dispense/slides, versione 1.1 Marzo 2021, distribuite dal docente durante lo svolgimento del corso, sotto licenza Creative Commons 4.0 CC BY-NC-ND.
Metodi didattici	Lezioni frontali con slides, ed esercitazioni di laboratorio mediante software open source (R + RStudio/Tidyverse).
Metodi di valutazione (indicare almeno la tipologia scritto, orale, altro)	<p>La prova d'esame consiste in:</p> <ol style="list-style-type: none"> 1. un progetto sull'analisi teorica di un algoritmo di Machine Learning o di un particolare aspetto teorico, con successiva una applicazione ad un problema reale utilizzando R, oppure R+Greta o anche R+Stan (questi ultimi due nel caso abbiate deciso di sviluppare un progetto in ambito Bayesiano); 2. una prova orale a distanza sulla piattaforma Teams, svolta a partire dal giorno e dall'ora dell'appello prescelto, nella quale verrà verificato il grado di effettiva conoscenza degli argomenti presentati. <p>La valutazione finale è in 30esimi ed è basata sui seguenti elementi:</p> <ul style="list-style-type: none"> • l'effettiva conoscenza degli argomenti presentati sul progetto, accertata attraverso la prova orale, è il presupposto necessario per non essere respinti

	<p>all'esame;</p> <ul style="list-style-type: none"> • in presenza del requisito imprescindibile di cui sopra, il 50% della valutazione per cento verrà attribuito sulla base dei contenuti effettivi del progetto, e il 50% sulla base della sua originalità e della chiarezza di presentazione.
<p>Criteria di valutazione (per ogni risultato di apprendimento atteso su indicato, descrivere cosa ci si aspetta lo studente conosca o sia in grado di fare e a quale livello al fine di dimostrare che un risultato di apprendimento è stato raggiunto e a quale livello)</p>	<p>Lo studente deve dimostrare una adeguata capacità di problem-solving.</p> <p>La presentazione del progetto individuale deve essere <u>originale</u> (altrimenti verrà considerata non valida), leggibile, completa e capace di generare interesse.</p> <p>Lo studente deve dimostrare di aver compreso adeguatamente l'argomento teorico esposto durante la verifica orale.</p>
<p>Altro</p>	<p>Ulteriori dettagli sulle modalità di svolgimento delle prove d'esame saranno forniti durante lo svolgimento delle lezioni frontali e di laboratorio.</p> <p>Nella stanza Teams dedicata all'insegnamento sarà disponibile un documento contenente una descrizione più dettagliata delle modalità di svolgimento delle prove d'esame.</p>