

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	24 febbraio 2020
Fine attività didattiche	29 maggio 2020

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, in un'ottica non supervisionata, per l'analisi dei dati testuali (data l'importanza fondamentale da questi ricoperta per l'analisi automatica dei dati testuali non strutturati ritrovati sul Web)

	<ul style="list-style-type: none"> • <i>Conoscenza e capacità di comprensione applicate</i> <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 dovrà essere in grado di utilizzare la versione 'moderna' basata su RStudio + Tidyverse</i> • <i>Autonomia di giudizio</i> <i>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).</i> • <i>Abilità comunicative</i> <i>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.</i> • <i>Capacità di apprendere</i> <i>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.</i>
<p>Contenuti di insegnamento</p>	<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 - Support Vector Machines - Modelli per dati testuali

<p>Programma</p>	<p>Teoria (32 h)</p> <ol style="list-style-type: none"> 1. Richiami di calcolo delle probabilità. 2. Inferenza frequentista. 3. Inferenza Bayesiana. Modelli gerarchici. Metodi computazionali: Gibbs sampling ed MCMC. 4. Regressione lineare semplice e multipla: inferenza frequentista e Bayesiana. 5. Regressione logistica semplice e multinomiale. Metodi numerici per la stima dei parametri: discesa del gradiente, IRLS. Cenni sull'inferenza Bayesiana. Interpretazione dei parametri: odds ratio. Classificazione logistica. Classificatori generativi e discriminativi. Analisi discriminante lineare e quadratica. Analisi discriminante lineare di Fisher. 6. Regolarizzazione: Ridge Regression, LASSO ed Elastic Net. Regressione non parametrica: basi polinomiali e splines, smoothing splines. Selezione automatica del parametro di
-------------------------	--

	<p>smoothing. Modelli additivi generalizzati (GAM).</p> <p>7. Apprendimento supervisionato e non-supervisionato. Teoria delle decisioni. Il teorema di trade-off bias-varianza. Inferenza per l'errore di generalizzazione: rischio empirico, rischio regolarizzato, rischio strutturale, la dimensione di Vapnik-Chervonenkis.</p> <p>8. Ancora sull'inferenza per l'errore di generalizzazione: Cp di Mallows, AIC e BIC, Cross-Validation, Bootstrap. La determinazione del modello ottimale in un'ottica Bayesiana. Metriche per la valutazione dell'accuratezza sul test set.</p> <p>9. Support Vector Machines. Regolarizzazione del margine e variabili slack. Funzioni Kernel. Classificazione a massimo margine e funzione di perdita hinge.</p> <p>10. Modelli per dati testuali I. Multinomial Bernoulli, Multinomial Naive Bayes.</p> <p>11. Modelli per dati testuali II. Cenni sull'inferenza variazionale. Latent Dirichlet Allocation (LDA).</p> <p>Laboratorio (30 h)</p> <ul style="list-style-type: none"> - Fondamenti di programmazione in R. - RStudio/Tidyverse. Le librerie di Tidyverse. - La libreria caret per l'automazione dei flussi di lavoro. - Interfacce verso Python/Anaconda e Keras/TensorFlow. - R Markdown.
Testi di riferimento	<p>Per la teoria:</p> <ul style="list-style-type: none"> - K. Murphy (2012): Machine Learning. A Probabilistic Perspective. The MIT Press, 1st Edition. ISBN-10/ASIN: 0262018020. ISBN-13: 978-0262018020. - 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>Per il laboratorio:</p> <ul style="list-style-type: none"> - G. Golemund, H. Wickam (2017). R for Data Science. O'Reilly Media, 1st Edition. ISBN-10/ASIN: 1491910399. ISBN-13: 978-1491910399 .
Note ai testi di riferimento	I testi di riferimento saranno supportati da un insieme di dispense/slides 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>Verifica orale Progetto individuale</p> <p>Durante lo svolgimento della prova d'esame lo studente dovrà esporre oralmente un argomento teorico estratto a caso tra quelli ricompresi nel programma, e presentare il</p>

	<p>progetto individuale discutendolo con il docente. Il progetto individuale deve essere preparato esclusivamente in R Markdown, e presentato in copia stampata (previo invio per per posta elettronica <u>in tempo utile</u>) nello stesso appello in cui si sostiene la verifica orale, oppure al più tardi (<u>senza alcuna possibilità di proroga</u>) entro i due appelli immediatamente successivi a quello nel quale è stata sostenuta la verifica orale. Gli studenti che non presentano il progetto individuale entro i due appelli immediatamente successivi a quello nel quale è stata superata la verifica orale, dovranno sostenere quest'ultima ex-novo.</p> <p>In ogni caso, la verifica orale deve essere sempre sostenuta come prima prova: <u>la presentazione del progetto è subordinata all'aver superato la verifica orale.</u></p> <p>Se durante la presentazione il progetto individuale viene giudicato non adeguato, lo stesso progetto potrà essere migliorato e ripresentato entro i due appelli immediatamente successivi (senza dover ripetere, in nessun caso, la prova orale). Se anche alla seconda discussione il progetto individuale viene valutato come non adeguato, il candidato dovrà ripetere ex-novo l'intera prova d'esame: nuova verifica orale e progetto individuale su un <u>nuovo</u> problema (seguendo le regole sopra esposte).</p> <p>Se la verifica orale viene superata, ad essa viene attribuito un voto in 30esimi. Se il progetto è considerato adeguato, ad esso viene attribuito un voto da 1 a 6. Il voto complessivo dell'esame viene attribuito sommando il voto della verifica orale e il voto attribuito al progetto individuale. Se tale somma risulta essere maggiore o uguale a 31, viene attribuita la votazione finale di 30 e lode.</p>
<p>Criteri 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>