UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

CONSIGLIO INTERCLASSE IN MATEMATICA

CORSO DI STUDIO LAUREA MAGISTRALE IN MATEMATICA (LM-40)

ANNO ACCADEMICO 2023-2024

INSEGNAMENTO STATISTICA PER IL MACHINE LEARNING

Principali informazioni sull'insegnamento	
Periodo di erogazione	Secondo semestre (26 febbraio 2024 – 31 maggio 2024)
Crediti formativi universitari (CFU)	7
Settore scientifico disciplinare (SSD)	MAT/06 – Probabilità e Statistica Matematica
Lingua di erogazione	Italiano
Modalità di frequenza	Facoltativa

Docente	
Nome e cognome	Marcello De Giosa
Indirizzo mail	marcello.degiosa@uniba.it
Telefono	+39 080 544 2707
Sede	Dipartimento di Matematica, piano 4, stanza 12.
Sede virtuale	
Pagina web	https://www.dm.uniba.it/it/members/degiosa
Ricevimento	Mercoledì, dalle 10 alle 12, con appuntamento per email.

Organizzazione della didattica				
	Totali	Didattica frontale	Pratica (esercitazioni)	Studio individuale
Ore	175	48	15	112
CFU	7	6	1	

Obiettivi formativi	
	Apprendere i principali metodi statistici predittivi Machine Learning e la loro corretta implementazione computazionale con il linguaggio di programmazione R e la collezione di suoi pacchetti tidymodels. Particolare attenzione è rivolta ai contributi teorici della Statistica ed alle buone pratiche metodologiche per produrre modelli statistici del Machine Learning di alta qualità e per interpretare e presentare in maniera corretta le conclusioni dell'analisi di strutture complesse di dati.

Prerequisiti	
	Calcolo, calcolo multivariato, algebra lineare, probabilità elementare, come offerti usualmente in un corso di studi in Matematica della classe L-35.

Syllabus	
Contenuti dell'insegnamento (Programma)	**Introduzione all' Apprendimento Statistico Automatico.** Compromesso tra accuratezza della previsione e interpretabilità del modello. Apprendimento supervisionato e non-supervisionato. Regressione contro classificazione. Compromesso tra distorsione e variabilità. Laboratorio introduttivo ad R. **Regressione Lineare (LR).** LR semplice e multipla. Stima dei coefficienti ed accuratezza. Predittori qualitativi. Confronto con KNN. Interazioni. Laboratorio R su regressione lineare. **Classificazione.** Regressione Logistica semplice e multipla, stima dei coefficienti, previsioni. Regressione logistica multinomiale. Metodi di

	classificazione generativi: analisi discriminante lineare e quadratica. Naive Bayes. Confronto di metodi di classificazione. Modelli lineari generalizzati: regressione Poisson. Laboratorio R su classificazione. **Metodi di ricampionamento**. Convalida incrociata e compromesso tra distorsione e variabilità. Convalida incrociata in classificazione. Il bootstrap. Laboratorio R su metodi di ricampionamento. **Regolarizzazione di modelli lineari.** Metodi di restringimento: Ridge e Lasso, messa a punto dei parametri. Metodi di riduzione della dimensione: Regressione alle componenti principali e minimi quadrati parziali. Dati di grandi dimensioni e relative problematiche. Laboratorio R su metodi di regolarizzazione. **Oltre la linearità.** Regressione polinomiale, funzioni a gradini, basi di funzioni. Regressione con splines e smoothing splines. Modelli additivi generalizzati in regressione e classificazione. Laboratorio R su metodi non lineari. **Metodi ad alberi.** Alberi decisionali di regressione e classificazione. Bagging, Random Forests, Boosting, alberi di regressione additivi bayesiani (BART). Laboratorio R su metodi ad albero. **Metodi a vettori di supporto (SVM).** Classificatore a massimo margine. Classificatore a vettori supporto. Support vector machines. Più di due classi: classificazioni uno-contro-uno e uno-contro-tutti. Relazioni con la regressione logistica. Laboratorio R su SVM. Curve ROC. Area sotto la curva. **Deep Learning.** Reti Neurali a strato singolo. Reti Neurali multistrato. Reti Neurali Convoluzionali: Convoluzione, raggruppamento, architetture, aumento dei dati, classificatori pre-addestrati. Propagazione all'indietro. Regolarizzazione e Discendente del Gradiente Stocastico. Abbandono. Messa a punto. Interpolazione e doppio discendente. Laboratorio R su apprendimento profondo con Keras e Torch. **Apprendimento non supervisionato.** Analisi delle componenti principali. Metodi di raggruppamento. K-means, raggruppamento Gerarchico. Laboratorio R su apprendimento non supervisionato. **Introduzione a tidymodels.** Rivisitazione degli argomenti trattati e dei laboratori R con tidymodels. **Rivisitazione dei metodi di ricampionamento.** Metodo di risostituzione. Convalida incrociata (CV). Convalida incrociata ripetuta (RCV). Convalida incrociata monte-carlo. Bootstrap. Stima della prestazione. Processare in parallelo. **Messa a punto del modello e dei parametri.** Ricerca su griglia e ricerca iterativa. Ottimizzazione bayesiana. Ottimizzazione con tempratura simulata. **Rivisitazione di metodi di riduzione della dimensione.** Analisi delle componenti principali (PCA), minimi quadrati parziali (PLS), analisi delle componenti indipendenti (ICA), approssimazione e proiezione su varietà uniformi (UMAP). Confronto di modelli basati su differenti tecniche di riduzione della dimensione.
Testi di riferimento	G. James, D. Witten, T. Hastie, R. Tibshirani – An introduction to Statistical Learning with application in R – 2023 – Springer T. Hastie, R. Tibshirani, J. Friedman – The Elements of Statistical Learning. Data Mining, Inference, and Prediction – 2009 – Springer K.P. Murphy – Probabilistic Machine Learning – 2022 – MIT Press J. Fan, R. Li, C. Zhang, H. Zou – Statistical Foundations of Data Science – 2020 – CRC Press

	M. Kuhn, k. Johnson – Applied Predictive Modeling – 2016 – Springer D.P. Kroese, Z.I. Botev, T. Taimre, R. Vaisman - Data Science and Machine Learning. Mathematical and Statistical Methods. - 2020 - CRC Press R. Irizarry – Introduction to Data Science – 2019 – CRC Press
Note ai testi di riferimento	
Materiali didattici	

Risultati di apprendimento previsti (secondo i Descrittori di Dublino)	
DD1 Conoscenza e capacità di comprensione	Acquisire i principali concetti e tecniche, e comprendere i codici R per l'Apprendimento Statistico Automatico.
DD2 Conoscenza e capacità di comprensione applicate	Essere capaci di applicare i risultati teorici ed i codici R per risolvere problemi reali dell'Apprendimento Statistico Automatico.
DD3-5 Competenze trasversali	*DD3 Autonomia di giudizio*: Essere in grado di scegliere i giusti strumenti e codici R dell'Apprendimento Statistico Automatico.
	DD4 Abilità comunicative: Essere in grado di esporre e spiegare chiaramente i metodi ed i risultati dell'Apprendimento Statistico Automatico.
	DD5 Capacità di apprendere: Essere in grado di applicare i metodi di apprendimento per migliorare la conoscenza dell'Apprendimento Statistico Automatico.

Metodi didattici
Lezioni frontali di persona con laboratorio R ed uso di diapositive.

Valutazione	
Modalità di verifica dell'apprendimento	L'esame finale si compone di una parte orale e di una parte di laboratorio R.
Criteri di valutazione	• *Conoscenza e capacità di comprensione*: padronanza e profonda comprensione dei metodi dell'Apprendimento Statistico Automatico e la loro traduzione in codici R. • *Conoscenza e capacità di comprensione applicate*: conoscenza dei metodi presentati e dei relativi codici e capacità di comprendere le differenze e peculiarità per le applicazioni a problemi reali. • *Autonomia di giudizio*: capacità di individuare autonomamente i metodi e codici più idonei a risolvere problemi reale. • *Abilità comunicative*: capacità di presentare i risultati, le conclusioni e le applicazioni di un metodo. • *Capacità di apprendere*: buona organizzazione delle conoscenza acquisite ed abilità nel trovare autonomamente sorgenti di approfondimento.
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	L'esame consiste nella esposizione e discussione di alcuni metodi tra quelli presentati a lezione e di alcuni codici R scelti a caso tra quelli presentati a lezione. Il voto, in trentesimi, dipenderà dalla padronanza degli argomenti e conoscenza dei codici, ma anche dalla capacità di esporre in maniera chiara.

Ulteriori informazioni	