Principali informazioni sull'insegnamento		
Denominazione dell'insegnamento	**Modellizzazione Statistica**	
Corso di studio	LM Data Science	
Anno Accademico	2023/24	
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS)		6 CFU
Settore Scientifico Disciplinare	SECS-S/01 (Statistica)	
Lingua di erogazione	Italiano	
Anno di corso	Primo	
Periodo di erogazione	2^ Semestre (01/03/24 – 07/06/24)	
Obbligo di frequenza	La frequenza è fortemente raccomandata	
Sito web del corso di studio	https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea	

Docente/i	
Nome e cognome	Massimo Bilancia
Indirizzo mail	massimo.bilancia AT uniba.it; massi.bilancia AT gmail.com
Telefono	–
Sede	Dipartimento di Informatica, Via Orabona 4, 70125, Bari.
Sede virtuale	https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/data-science/data-science
Sito web del docente	–
Ricevimento (giorni, orari e modalità, es. su appuntamento)	In qualsiasi giorno previo appuntamento concordato

Syllabus	
Obiettivi formativi	Il corso di Modellizzazione Statistica mira a fornire allo studente del corso di LM in Data Science le conoscenze teoriche essenziali per applicare in modo efficiente le tecniche di Machine Learning che vengono apprese durante il corso. Oggetto dell'insegnamento saranno, pertanto, i fondamenti teorici dei due principali paradigmi

	inferenziali (frequentista e Bayesiano), la conoscenza dei principali metodi ottimizzazione e Monte Carlo per la stima dei parametri e la determinazione del modello ottimale, la teoria dell'apprendimento statistico supervisionato e della generalizzazione. In parallelo, durante le lezioni di laboratorio, gli studenti apprenderanno i fondamenti della programmazione in R, che costituisce la seconda piattaforma (in termini di numero di utenti) in ambito Data Science dopo Python.
Prerequisiti	**Matematica**: vettore gradiente, punti di massimo e di minimo delle funzioni di più variabili, estremi vincolati e teorema dei moltiplicatori di Lagrange (cfr: Fondamenti di Matematica per la Data Science) **Elementi di Programmazione**: principali strutture dati, elementi di base della programmazione orientata agli oggetti, in particolare in Python (cfr: Fondamenti di Programmazione per la Data Science) **Fondamenti di Statistica**: gli elementi di base impartiti attraverso un qualunque insegnamento di Statistica erogato in una laurea Triennale.
Contenuti di insegnamento (Programma)	**Teoria**: 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 valutazione dell'accuratezza sul test set. **Laboratorio**: Fondamenti di programmazione in R – RStudio/Tidyverse – Le librerie di Tidyverse - La libreria caret per l'automazione dei flussi di lavoro – R Markdown e Quarto per la reportistica – Stan e Greta per l'inferenza Bayesiana in R.
Testi di riferimento	1. James G., Witten D., Hastie T., Tibshirani R. (2021) *An Introduction to Statistical learning*, 2nd Edition, Springer Nature. ISBN-10/ASIN: ISBN-10: 8829930946. ISBN–13: 978-1071614174. Liberamente scaricabile al seguente indirizzo: https://www.statlearning.com/ 2. Murphy K.P. (2022) Probabilistic Machine Learning: An Introduction. The MIT Press. ISBN–13: 978-0262046824. Liberamente scaricabile all'indirizzo https://probml.github.io/pml-book/book1.html 3. G. Grolemund, H. Wickam (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2nd Edition, O'Reilly Media. ISBN–13: 978-1492097402 4. Dispense fornite dal docente, versione 1.4 Marzo 2024 distribuite sotto licenza Creative Commons 4.0 CC BY-NC-ND
Note ai testi di riferimento	1. Può essere utilizzato come testo di riferimento/consultazione fino all'argomento "Modelli lineari generalizzati (GAM)" della parte teorica. 2. Può essere utilizzato come testo di riferimento/consultazione a partire dall'argomento "apprendimento supervisionato e non-supervisionato" della parte teorica 3. Può essere utilizzato come testo di riferimento per la parte di laboratorio. 4. Contengono il core del materiale di studio.

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Laboratorio	Studio individuale
150 ore	32 ore	30 ore	88 ore
CFU/ETCS			
6 CFU	4 CFU (1 CFU = 8 ore)	2 CFU (1 CFU = 15 ore)	

Metodi didattici	
	• Per la teoria: didattica frontale in presenza (32 ore) • Per il laboratorio: esercitazioni pratiche in laboratorio utilizzando R + RStudio/Tidyverse (30 ore)

Risultati di apprendimento previsti	
Conoscenza e capacità di comprensione	• Acquisizione di competenze avanzate nell'ambito dell'inferenza statistica classica, dell'inferenza statistica Bayesiana e della teoria dell'apprendimento statistico, anche attraverso la lettura di testi avanzati sull'argomento ed articoli di ricerca
Conoscenza e capacità di comprensione applicate	• Capacità di applicare gli strumenti teorici appresi a situazioni reali, senza ricadere in una applicazione acritica degli algoritmi di Machine Learning all'interno del flusso di lavoro tipico della Data Science
Competenze trasversali	**Autonomia di giudizio** • Acquisizione di conoscenze teoriche e pratiche utili a poter formulare giudizi autonomi sulla qualità dei risultati raggiunti nelle applicazioni pratiche, e sulla loro generalizzabilità **Abilità comunicative** • Sviluppo di abilità avanzate di reportistica dinamica attraverso l'utilizzo di R/R Studio/Tidyverse/Quarto **Capacità di apprendere in modo autonomo** • Capacità di selezionare e leggere la letteratura riguardanti gli sviluppi più recenti nel campo della teoria dell'apprendimento statistico, in particolare con applicazioni alle capacità di generalizzazione dei modelli di Deep Learning

Valutazione	
Modalità di verifica dell'apprendimento	Le prove di esame si compongono di due distinte sotto-prove: • **laboratorio**: durata 1h e 30m, consistente nella produzione di un report mediante R/R Studio/Quarto su quanto viene richiesto nella traccia. • **scritto**: durata 1h e 30m, consistente nella produzione di un tema scritto su due quesiti riguardanti gli argomenti svolti a lezioni durante la parte teorica

	Alla prima iscrizione, le due prove devono essere svolte nello stesso appello al quale ci si è iscritti. La prova di laboratorio viene valutata come SI/NO, alla prova scritta verrà attribuito un voto in trentesimi che diventa il voto finale verbalizzato in caso di accettazione. Le regole sono: 1. non superare la prova di laboratorio comporta la ripetizione dell'intero esame 2. non superare la prova scritta comporta la sola ripetizione della prova scritta in un appello successivo 3. se la prova di laboratorio è superata, è possibile rifiutare selettivamente la sola prova scritta (mantenendo il laboratorio) e ripeterla in un appello successivo
Criteri di valutazione	• **Conoscenza e capacità di comprensione:** Esposizione critica dei concetti teorici appresi durante il corso • **Conoscenza e capacità di comprensione applicata** Capacità di sviluppare in maniera rapida ed efficiente il codice in R per prototipizzare un problema • **Autonomia di giudizio** Capacità di selezionare autonomamente gli strumenti teorici più adeguati al trattamento algorimo della tipologia di dati analizzati • **Abilità comunicative:** Capacità di utilizzare strumenti di reportistica avanzata in R/R Studio/Quarto per la presentazione dei risultati • **Capacità di apprendere:** Capacità di utilizzare in modo critico anche la letteratura più recente
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Il voto finale in trentesimi sarà attribuito sulla base di una media non ponderata dei voti attribuiti ai criteri di valutazione descritti sopra
Altro	Si suggerisce agli studenti di affidarsi esclusivamente alle informazioni/comunicazioni fornite sui siti ufficiali del Dipartimento di Informatica, ovvero sui gruppi social solo se costituiti e amministrati esclusivamente dai docenti dei relativi insegnamenti. I programmi degli insegnamenti sono disponibili qui: https://elearning.uniba.it/course/index.php?categoryid=283 I regolamenti didattici e i Manifesti degli Studi dei CdS sono disponibili qui: https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-dilaurea/corsi-di-laurea _____ Tutto il materiale del corso e le informazioni aggiuntive saranno distribuiti in tempo reale sulla piattaforma didattica di UniBA raggiungibile all'indirizzo: https://elearning.uniba.it/ La chiave di iscrizione alla pagina dedicata al corso di Modellizzazione Statistica è: msds2324

Main information on the course

Course name	**Statistical Modelling**	
Degree	LM Data Science	
Academic year	2023/24	
European Credit Transfer and Accumulation System (ECTS), in Italian Crediti Formativi Universitari (CFU)		6 CFU (4 T1 + 2 T2) (each CFU corresponds to 25 hours (h) of student's time); CFU are of type T1, T2 or T3 T1 = 8 h lecture + 17 h individual study T2 = 15 h practice + 10 h individual study T3 = 25 h individual study
Settore Scientifico Disciplinare	SECS-S/01 (Statistics)	
Course language	Italian (English on demand)	
Anno di corso	First	
Periodo di erogazione	Second semester	
Obbligo di frequenza	It is highly recommended to attend classes	
Sito web del corso di studio	https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea	

Teacher(s)

Name and Surname	Massimo Bilancia
email	massimo.bilancia AT uniba.it or massi.bilancia AT gmail.com
phone	--
office	Department of Computer Science, Via Orabona 4, 70125, Bari.
e-learning platform	https://elearning.uniba.it/ (registration key: msds2324)
Teacher's homepage	--
Office hours	Every day by prior appointment

Syllabus

Course goals	The Statistical Modeling course aims to provide students of the LM in Data Science course with the essential theoretical knowledge to efficiently apply the Machine Learning techniques learned during the course. The course therefore covers the theoretical foundations of the two main inference paradigms (frequentist and Bayesian), knowledge of the main optimization and Monte Carlo methods for parameter estimation and optimal model determination, the theory of supervised statistical learning and generalization. In parallel, students learn the basics of programming in R in laboratory classes, which is the second most important platform (in terms of the number of users) in the field of Data Science after Python.
Prerequisites/requirements	**Mathematics**: gradient vector, maximum and minimum points of functions of several variables, bounded extrema and Lagrange's multiplier theorem (see: Fundamentals of Mathematics for Data Science) **Elements of programming**: most important data structures, basic elements of object-oriented programming, especially in Python (see: Fundamentals of Programming for Data Science) **Fundamentals of Statistics**: the basic elements taught in every statistics course in a three-year degree program.

Course program	**Theory**: Overview of probability theory – Monte Carlo methods – Frequentist inference and maximum likelihood estimation: Properties in the frequentist sense and numerical methods – Bayesian inference and hierarchical models – Computational methods for Bayesian inference: Gibbs sampling and Markov Chain Monte Carlo (MCMC) – Classification regression models – Simple and multiple linear regression: Frequentist and Bayesian inference – Simple and multinomial logistic regression – Numerical methods for parameter estimation: gradient descent, IRLS and notes on Bayesian inference – Interpretation of parameters: odds ratio – Logistic classification – Linear, quadratic and kNN discriminant analysis – Generative and discriminative classifiers – - Nonparametric regression: polynomial bases and splines, smoothing splines – Automatic selection of the smoothing parameter – Generalized additive models (GAM) – Supervised and unsupervised learning – Decision theory – The bias-variance trade-off theorem for regression models – The bias-variance trade-off in classification problems – Inference for the generalization error: Mallows' Cp, AIC and BIC, cross-validation, bootstrap – Regularization: ridge regression, LASSO and Elastic Net – Determining the optimal model from a Bayesian perspective – The PAC framework and the Vapnik-Chervonenkis dimension – Metrics for evaluating accuracy on the test set. **Laboratory**: basics of programming in R – RStudio/Tidyverse – The Tidyverse libraries - The caret library for workflow automation – R Markdown and Quarto for reporting – Stan and Greta for Bayesian inference in R.
Books of reference	1. James G., Witten D., Hastie T., Tibshirani R. (2021) *An Introduction to Statistical learning*, 2nd Edition, Springer Nature. ISBN-10/ASIN: ISBN-10: 8829930946. ISBN–13: 978-1071614174. Liberamente scaricabile al seguente indirizzo: https://www.statlearning.com/ 2. Murphy K.P. (2022) Probabilistic Machine Learning: An Introduction. The MIT Press. ISBN–13: 978-0262046824. Liberamente scaricabile all'indirizzo https://probml.github.io/pml-book/book1.html 3. G. Grolemund, H. Wickam (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2nd Edition, O'Reilly Media. ISBN–13: 978-1492097402 4. Teaching slides, version 1.4 Marzo 2024 distributed under Creative Commons 4.0 CC BY-NC-ND
Notes to the books	1. It can be used as a reference/consulting text up to the topic "Generalized Linear Models (GAM)" of the theoretical part. 2. It can be used as a reference/consulting text from the topic "Supervised and Unsupervised Learning" of the theoretical part onwards 3. It can be used as a reference text for the laboratory part 4. They contain the core study material.

Organization of the didactic activities

Hours			
Total	Lectures	Practice sessions	Individual study
hours 150	hours 32	hours 30	hours 88
CFU/ETCS			
CFU 6	CFU 4 (1 CFU = 8 hours)	CFU 2 (1 CFU = 15 hours)	

Teaching methods

Theory: frontal teaching (32 hours)
Laboratory: practical exercises in the laboratory with R + RStudio/Tidyverse (30 hours)

Expected learning outcomes

Knowledge and understanding	• Acquisition of advanced knowledge in the field of classical statistical inference, Bayesian statistical inference and statistical learning theory, also by reading advanced texts on the topic and research articles
Applying knowledge and understanding	• Ability to apply the theoretical tools learned to real-life situations without falling into an uncritical application of Machine Learning algorithms within the typical data science workflow
Other skills	*Making judgements* Acquisition of theoretical and practical knowledge useful to formulate an independent judgment on the quality of results obtained in practical applications and their generalizability *Communication* Developing advanced dynamic reporting skills using R/R Studio/Tidyverse/Quarto *Learning skills* - Ability to select and read literature on the latest developments in the field of statistical learning theory, especially regarding the generalizability of Deep Learning models

Assessment	
Assessment methods	The examination tests consist of two different subtests: • **Laboratory**: duration 1 hour and 30 minutes, consisting of the preparation of a report with R/R Studio/Quarto on a practical data analysis problem • **Written exam**: lasting 1 hour and 30 minutes, consisting of the preparation of a written essay on two questions on the topics covered in the theoretical part If you are registering for the first time, both tests must be taken in the same session for which you have registered. The lab test will be graded YES/NO and the written test will be graded out of thirty points, which will be the final grade if the test is accepted. The rules are: 1. if you fail the lab test, the entire exam will be repeated 2. if you fail the written test, you only must repeat the written test in a later examination session 3. if the lab test is passed, it is possible to reject only the written test (keep the lab test) and retake it in a later session
Evaluation criteria	**Knowledge and understanding:** Critical examination of the theoretical concepts learned in the course **Applied knowledge and understanding:** Ability to quickly and efficiently develop code in R to prototype a problem **Independent judgment:** Ability to independently select the most appropriate theoretical tools for the algorithmic treatment of the analyzed data type **Communication skills:** Ability to use advanced reporting tools in R/R Studio/Quarto to present results **Ability to learn:** Ability to critically utilize even the latest scholarly literature
Measurements and final grade	The final mark out of thirty points is awarded based on an unweighted average of the marks for the assessment criteria described above

Further information

Students are advised to rely solely on the information/communications on the official Computer Science Department website, or on social groups if these are set up and managed solely by the lecturers of the relevant courses.

Course syllabi are available here:
https://elearning.uniba.it/course/index.php?categoryid=283

Course syllabi and course manifestos are available here:
https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-dilaurea/corsi-di-laurea

All course materials and additional information will be distributed in real time on the UniBA learning platform, which can be accessed at:
https://elearning.uniba.it/

The registration key on the page for the Statistical Modeling course is:
msds2324