UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

CONSIGLIO INTERCLASSE IN MATEMATICA

COURSE OF STUDY TWO-YEAR MASTER OF SCIENCE PROGRAMME IN MATHEMATICS

ACADEMIC YEAR 2023-2024

ACADEMIC SUBJECT STATISTICS FOR MACHINE LEARNING

General information	
Term	Second semester (February 26, 2024 – May 31, 2024)
European Credit Transfer and Accumulation System credits (ECTS)	7
SSD	MAT/06 – Probability and Mathematical Statistics
Language	Italian
Mode of attendance	Not mandatory

Lecturer	
Name and surname	Marcello De Giosa
E-mail	marcello.degiosa@uniba.it
Telephone	+39 080 544 2707
Department and office	Department of Mathematics, floor 4, room 12.
Virtual meeting room	
Web page	https://www.dm.uniba.it/it/members/degiosa
Office hours	Wednesday, from 10 a.m. to 12 a.m., by email appointment.

Work schedule	Total	Lectures	Hands-on learning (recitations)	Self-study
Hours	175	48	15	112
ECTS credits	7	6	1	

Learning objectives	
	Acquiring knowledge of the main statistical predictive methods of Statistical Machine Learning, and their correct computational implementation with the R programming language e the package collection tidymodels. Prticular attention is paid to the theoretical contributions of Statistics and to good methodological practices to produce high quality statistical models of Machine Learning and to correctly interpret and present the conclusions of the analysis of complex data structures.

Course prerequisites	
	Calculus, multivariate calculus, linear algebra, elementary probability, as usually offered during a degree in Mathematics of L-35 class.

Syllabus	
Course contents	**Introduction to Statistical Machine Learning.** Trade-off between prediction accuracy and model interpretability. Supervised and unsupervised learning. Regression versus classification. Trade-off between bias and variance. Introductory R lab. **Linear regression (LR).** Simple and multiple LR. Coefficients estimation and accuracy. Qualitative predictors. Comparison with KNN. Interactions. R lab on

	LR. **Classification.** Logistic Regression. The simple and multiple cases. Generative classification methods: linear and quadratic discriminant analysis. Naïve Bayes. Comparing classification methods. Generalized linear models: Poisson regression. R lab on classification. **Resampling methods.** Cross-validation and bias-variance tradeoff. Cross validation in classification. The bootstrap. R lab on resampling. **Linear model regularization.** Shrinkage: Ridge and Lasso. Hyperparameters tuning. Dimension reduction: principal component regression and partialleast squares. High dimensional data. R lab on regularization. **Beyond linearity.** Polynomial regression, step functions, basis functions. Splines regression and smoothing splines. Generalize additive models for regression and classification. R lab on nonlinearity. **Decision tree.** Regression and classification regression trees. Bagging, Random Forests, Boosting, Bayesian additive regression trees (BART). R lab on decision tree methods. **Support vector machines.** Maximal margin classifier. Support vectors classifier. Support vector machines. More than two classes: one-versus-one and one-versus-all methods. Comparing with logistic regression. R lab on support vector machines. **Deep Learning.** Single Layer Neural Networks. Multilayer Neural Networks. Convolutional Neural Networks: Convolution layer, pooling layer, architecture, data augmentation, pretrained classifier. Backpropagation. Regularization and Stochastic Gradient Descent. Dropout. Tuning. Interpolation and Double Descent. R lab on deep learning with Keras and Torch. **Unsupervised Learning.** Principal Component Analysis (PCA). Clustering: K-means, hierarchical clustering. R lab on unsupervised learning. **Introduction to tidymodels.** Review of R lab with tidymodels. **Review of resampling methods.** Resubstitution. Cross-validation. Repeated Cross-validation. Monte-Carlo Cross-validation. Bootstrap. Performance estimation. Parallel processing. **Model and parameters tuning.** Grid search and iterative search. Bayesian optimization. Simulated annealing. **Review of dimension reduction methods.** Principal Components Analysis (PCA), Partial Least Squares (PLS), Independent Component Analysis (ICA), Uniform Manifolds Approximation and Projection (UMAP). Comparison of models based on different dimension reduction methods. R lab on dimensions reduction.
Reference books	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
Additional course materials	
Repository	

Expected learning outcomes	
Knowledge and understanding	Acquiring the main concepts and techniques, and R code understanding for Statistical Machine Learning.
Applying knowledge and understanding	Being able to apply the theoretic results and R code to solve real world Statistical Machine Learning problems.
Soft skills	*Making judgements*: Being able to choose the right Statistical Machine Learning tools and R code.
	Communication skills: Being able to report and clearly explain the Statistical Machine Learning methods and results.
	Learning skills: Being able to apply learning methods to improve knowledge in the field of Statistical Machine Learning.

Teaching methods	
	In person frontal theoretical lessons and R laboratory with use of slides.

Assessment	
Assessment methods	The final exam consists of an oral part and an R laboratory part.
Evaluation criteria	• *Knowledge and understanding*: mastering and deep understanding of the Statistical Machine Learning methods and their translation in modern R code. • *Applying knowledge and understanding*: being able to apply the Statistical Machine Learning methods and related R code to solve practical real world problems. • *Making judgement*: mastering the ability to choose the right Statistical Machine Learning methods and R code to solve real problems. • *Communication skills*: ability to explain the applied methods and their results in a clear way. • *Learning skills*: well organizing the acquired knowledge and being able to autonomously find sources of insight.
Grading policy	The exam consists in the discussion of some methods among those proposed during the classes and the discussion of some R code chosen at random from those presented during the classes. The vote, out of thirty, will depend on the mastery of the topics, knowledge of the codes, but also on the ability to clearly explain results.

Further information	