UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Scienze Economiche, Aziendali e Statistiche
ACADEMIC YEAR	2023/2024
MASTER'S DEGREE (MSC)	STATISTICS AND DATA SCIENCE
SUBJECT	STATISTICAL MACHINE LEARNING
TYPE OF EDUCATIONAL ACTIVITY	C
AMBIT	21031-Attività formative affini o integrative
CODE	23500
SCIENTIFIC SECTOR(S)	SECS-S/01
HEAD PROFESSOR(S)	SOTTILE GIANLUCA Ricercatore a tempo determinato Univ. di PALERMO
OTHER PROFESSOR(S)	
CREDITS	6
INDIVIDUAL STUDY (Hrs)	108
COURSE ACTIVITY (Hrs)	42
PROPAEDEUTICAL SUBJECTS	
MUTUALIZATION	
YEAR	2
TERM (SEMESTER)	1° semester
ATTENDANCE	Not mandatory
EVALUATION	Out of 30
TEACHER OFFICE HOURS	**SOTTILE GIANLUCA** Monday 10:00 12:00 Ufficio del docente Wednesday 10:00 12:00 Ufficio del docente

DOCENTE: Prof. GIANLUCA SOTTILE

PREREQUISITES	The student should have a basic understanding of the processes involved in data analysis, outliers or missing data, and the representation and interpretation of data in a multidimensional space. Knowledge of R software is required.
LEARNING OUTCOMES	Knowledge and understanding. Knowledge of classification, clustering and forecasting methods. Acquisition of the language and terminology of the discipline. Ability to choose the most appropriate models for each problem and evaluate them with objective metrics. Applying knowledge and understanding. The student must be able to: 1. correctly formulate a problem 2. use the R statistical environment to apply the different types of learning algorithms (even in combination among them) to solve each problem 3. interpret the results in a multidisciplinary environment and translate them into operational decisions. Making judgments Critically understand machine learning methods' characteristics, potential and limitations for solving specific problems. Communication skills The student must be able to: 1. communicate the concepts and data analysis techniques studied to both specialists and non-specialists in the subject 2. justify the choices of tools used for the analysis 3. communicate the results in an appropriate language. Learning skills The student will have developed the learning skills necessary to deepen the study of the most recent machine learning techniques, even in software environments other than the one used in the classroom.
ASSESSMENT METHODS	Presentation of two reports (belonging to the "gamification" phase) and oral exam. The verification of learning takes place through: - For those who take the exam during the session following the course delivery session 1. Delivery of an intermediate report containing the analysis of a dataset assigned about halfway through the course to groups of students of at most three people. The delivery and presentation of the analysis results by a representative for each group is expected after the mid-term break. 2. Delivery of a final report containing the analysis of a dataset assigned at the end of the course (delivered at least one week before the oral exam) and its presentation (in PowerPoint or other). - For those who take the exam in other sessions 1. Delivery of an intermediate report containing the analysis of an assigned dataset at least three weeks before the start of the session. Delivery is expected one week before the first appeal and the presentation of the analysis results during one of the appeals. 2. Delivery of a final report containing the analysis of an assigned dataset at least three weeks before the start of the session (delivered at least one week before the oral test) and its presentation (in PowerPoint or other). The final and the intermediate report (at most eight pages, plus any graphs and tables) must be delivered, accompanied by the corresponding R codes. The oral test will consist of the presentation (in PowerPoint or another) of the analysis and the results described in the final report to evaluate better the student's knowledge and skills and ability to transmit them with suitable statistical language. FINAL EVALUATION The final assessment of the exam will take into consideration three aspects: i) mastery of the topics, ii) the ability to apply knowledge and iii) the property of language, evaluated overall in the intermediate (weight 0.3) and final (weight 0.5) reports and the oral exam (weight 0.2). The evaluation will be sufficient if the candidate has chosen the appropriate analysis methods for both reports, even if the analysis will not be efficient. Furthermore, the candidate must demonstrate sufficient argumentative and expository ability during the oral exam for a sufficient evaluation. The more the examinee gives evidence, in writing the reports and in the oral exam, of his argumentative and expository skills, statistical language properties, and efficient

	use of the R programming language for data analysis, the more the evaluation will be positive. The Selection Committee is formed by the class chair and at least one other teacher, professor, assistant professor, or expert.
EDUCATIONAL OBJECTIVES	This course represents a point of contact between non-linear models and the corresponding learning algorithms to solve problems of the following types: - Classification; - Grouping; - Regressions; - Prediction. Each of these problems requires an approach that can be different in learning. Thus, the course begins by reviewing the basic concepts and definitions of the different types of machine learning: supervised, unsupervised, semi-supervised, active, and reinforcement learning. Next, we will learn how to evaluate the results of these problems, the different metrics that exist, the need to partition datasets to ensure acceptable performance and possible improvements that may arise, such as boosting or bagging techniques to generate ensembles. Eventually, students can develop the most appropriate machine learning model, combining different types and interpreting the results of the provided solution in a multidisciplinary environment.
TEACHING METHODS	The course will be held in English. The course will be divided into lectures and exercises. All the theoretical topics developed in the frontal lessons will be addressed in application terms through computer-statistical laboratory activities (individual or group) using the R programming environment. During the course, a "gamification" phase will be carried out (belonging to the group of innovative teaching methodologies), i.e. the application of game design mechanics to stimulate learning.
SUGGESTED BIBLIOGRAPHY	Lecture notes made available by the professor on the University portal. Online resources indicated by the teacher during the course. - A. Ghatak (2017). Machine Learning with R, Springer (Chap. 1, 3, 4, 5) - S.V. Burger (2018). Introduction to Machine Learning with R, Publisher(s): O'Reilly Media, Inc. (Chap. 2, 3, 4, 5, 6, 7, 8) - E. Alpayidin, F. Bach (2014). Introduction to Machine Learning, Fourth Edition, The MIT Press (Chap. 1, 2, 6, 7, 9, 11, 13, 18) - B. Boehmke, B.M. Greenwell (2020). Hands-On Machine Learning with R, First Edition, CRC Press Taylor & Francis Group (Chap. 1, 2, 4, 5, 8, 9, 10, 11, 12, 13, 14, 17, 20, 21) - C. Lesmeister, S.K. Chinnamgari (2019). Advanced Machine Learning with R, Packt Publishing Limited (Chap. 2, 3, 5, 6, 7, 8, 9, 10) - A. Kassambara (2017). Machine Learning Essentials: Practical Guide in R. STHDA (Chap. 3, 5, 12, 13, 14, 21, 24, 26, 27, 28, 29, 31, 32, , 33, 34, 35)

SYLLABUS

Hrs	Frontal teaching
2	Preliminary topics - Definitions: sample, model, variable, algorithm, dimensionality, ... - Standardization and coding - Variables selection - Variables extraction: Decomposition into singular values (SVD), Principal Component Analysis (PCA), ...
4	Approaches to learning and related problems - Supervised learning: classification problems; and regression and prediction problems - Unsupervised Learning: K-Means and hierarchical clustering - Semi-supervised approaches: semi-supervised and active learning - Reinforcement learning: optimisation problems
4	Model assessment - Overtraining and overfitting - Data set splitting: K-fold and leave-one-out - Assessing the performance of machine learning models: classification and regression problems - Model improvement: boosting, ensembles, and bagging
2	Support Vector Machines - Introduction - Optimum separation hyperplane - The kernel trick - Supported Vector Regressor (SVR)

SYLLABUS

Hrs	Frontal teaching
2	Decision trees - Representation - Entropy and information gain - Pruning - Classification and Regression Trees (CART) - Random Forest (RF)
10	Current trends in Machine Learning - Naive Bayes - Gradient Boosting - K-Nearest Neighbours (KNN) - Artificial Neural Networks (ANN) - Deep Learning.

Hrs	Practice
18	Implementation of the models described in the theoretical lessons: - Preprocessing of data sets: standardisation, coding, variables selection and extraction; - Model assessment: K-fold, Leave-one-out, ...; - Supervised learning: classification and regression problems (e.g., linear, logistic, support vector machine, decision tree-based, random forest, naive Bayes, gradient boosting ...); - Unsupervised learning: clustering problems; - Semi-supervised learning: active and reinforcement learning ; - Deep learning and artificial neural network.