Principali informazioni sull'insegnamento	
Anno di corso	2023/2024
Periodo di erogazione	*marzo 2024 – maggio 2024*
Crediti formativi universitari (CFU/ETCS):	*6*
SSD	*ING-INF/05*
Lingua di erogazione	*Italiano o Inglese*
Modalità di frequenza	*Frequenza facoltativa*

Docente	
Nome e cognome	*Dott. Roberto Cilli*
Indirizzo mail	*roberto.cilli@uniba.it*
Telefono	*327 730 8761*
Sede	*Dipartimento di Fisica, Primo Piano, Stanza 124*
Sede virtuale	*Canale Teams*
Ricevimento	*Ogni venerdì – ore 15-19*

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
150	*32*	*30*	*88*
CFU/ETCS			
6	*4*	*2*	

Obiettivi formativi	*Il corso di insegnamento in machine learning ed intelligenza artificiale ha come obiettivo formativo quello di fornire agli studenti le competenze necessarie per comprendere e implementare algoritmi di apprendimento automatico per l'estrazione di conoscenza in grandi moli di dati multi-omici.*
Prerequisiti	*Non vi sono prerequisiti specifici differenti da quelli richiesti per l'accesso al corso di laurea. Tuttavia, conoscenze preliminari di programmazione in linguaggio R e di statistica possono aiutare ad affrontare i contenuti previsti dall'insegnamento.*

Metodi didattici	*La modalità di erogazione dell'insegnamento prevede delle lezioni frontali alternati ad esercitazioni in aula. L'erogazione dell'insegnamento segue la filosofia del "learning-while-doing" ponendo l'enfasi sull'applicazione degli strumenti e della teoria dell'apprendimento a casi di studio reali e di interesse per Bioinformatici e Scienziati dei dati.* *Le lezioni frontali fanno affidamento su supporti multimediali (presentazioni PowerPoint) ed hanno come obiettivo l'insegnamento dei fondamenti della teoria dell'apprendimento, delle tecniche di apprendimento supervisionato e delle procedure di validazione dei modelli data-driven.* *Le esercitazioni in aula prevedono la scrittura di codice in linguaggio R in grado di eseguire tutte le procedure utili per pulire i dati in ingresso, eseguire analisi preliminari, addestrare e validare algoritmi di apprendimento.*

Risultati di apprendimento previsti ***Da indicare per ciascun Descrittore di Dublino (DD=*** ***DD1* Conoscenza e capacità di comprensione** ***DD2* Conoscenza e capacità di comprensione applicate** ***DD3-5* Competenze trasversali**	- ***Descrittore di Dublino 1****: conoscenza e capacità di comprensione;* ○ Comprensione dei teoremi generali della teoria dell'apprendimento supervisionate e delle loro implicazioni; ○ Comprensione delle specificità degli algoritmi di apprendimento automatico introdotte nel corso - ***Descrittore di Dublino 2****: capacità di applicare conoscenza e comprensione;* ○ Capacità di applicare le procedure di analisi illustrate durante il corso per condurre analisi su dati omici mai visti prima. - ***Descrittore di Dublino 3****: capacità critiche e di giudizio (occorre indicare le attività che concorrono allo sviluppo di tali abilità.* • *Autonomia di giudizio* *Al termine dell'insegnamento lo/la studente/studentessa dovrà essere in grado di* • Capacità di applicare le procedure di analisi illustrate durante il corso per condurre analisi su dati omici mai visti prima. • Capacità di formulare giudizio autonomo circa le performance di classificazione misurate per un certo problema con riferimento a limiti o margini di miglioramento dell'algoritmo di apprendimento proposto • Comprendere implicazioni etiche delle varie fonti di bias nella ricerca accademica. - ***Descrittore di Dublino 4****: capacità di comunicare quanto si è appreso.* *Abilità comunicative* *Al termine dell'insegnamento lo/la studente/studentessa dovrà essere in grado di* • Realizzare grafici efficaci nel comunicare esistenza di relazioni associative • Comunicare i risultati scientifici raggiunti ai propri interlocutori con l'ausilio di supporti quali diapositive - ***Descrittore di Dublino 5****: capacità di proseguire lo studio in modo autonomo nel corso della vita. Capacità di apprendere in modo autonomo:* *Al termine dell'insegnamento lo/la studente/studentessa dovrà essere in grado di* • Capacità di applicare le procedure di analisi illustrate durante il corso per condurre analisi su dati omici mai visti prima.

Contenuti di insegnamento (Programma)	*• Introduction: what is Machine Learning; Problems, data, and tools.* *• An introduction to R.* *• Feature Engineering.* *• Dimensionality reduction; latent space methods; Principal Component Analysis (PCA);* *• Regression problem: linear regression, gradient descent.* *• Classification problems: Overfitting and complexity; bias and variance; training, validation, and test; Cross validation procedure; confusion matrix, accuracy, precision, recall, f1 score.* *• Linear classifiers; Naive Bayes model; Neural Networks, Support Vector Machine.* *• Decision Tree; Ensemble: Bagging, Random forests, Boruta, VSURF.* *• Unsupervised learning: clustering, k-means, hierarchical agglomeration.* *• Hints on Explainable Artificial Intelligence (XAI): SHAP Model.* *• Hints on CNNs and Unet-128.* *• Hints on Cohort Studies.*
Testi di riferimento	***Christopher M. Bishop, Pattern Recognition and Machine Learning***
Note ai testi di riferimento	
Materiali didattici	*Presentazioni PowerPoint su canale Teams.*

Valutazione	
Modalità di verifica dell'apprendimento	*La verifica dell'apprendimento prevede la presentazione di un progetto di analisi di un dataset multi-omico adeguatamente descritto da articolo internazionale revisionato tra pari. Il progetto prevede tre fasi:* • *sviluppo software e addestramento di algoritmi* • *misura e validazione delle performance di classificazione/regressione* • *realizzazione di un PowerPoint e colloquio orale di 20 minuti (10 slide) relativamente alle performance della catena di analisi proposta* *Il colloquio orale ha la finalità di comprendere se lo studente ha acquisito la capacità di applicare autonomamente gli strumenti descritti e dimostrati in aula a problemi scientifici aperti.* *Tali competenze verranno comprovate durante il colloquio finale con l'ausilio di una presentazione PowerPoint. La presentazione PowerPoint prevede alcune slide introduttive sul problema scientifico oggetto di indagine, sulla catena di analisi e sui metodi impiegati, ed infine, alcune slide conclusive relativamente a risultati ottenuti e limiti delle procedure adottate rispetto al particolare dato investigato.* *In particolare, la modalità di valutazione dell'esame farà particolare attenzione a misurare la capacità dello studente di individuare relazioni di associazione statistica in dati di elevata dimensionalità.*

Criteri di valutazione	• *Conoscenza e capacità di comprensione:*
	  • Comprensione dei teoremi generali della teoria dell'apprendimento supervisionate;
	• *Conoscenza e capacità di comprensione applicate:*
	  • *Comprensione implicazione dei teoremi generali della teoria dell'apprendiemento supervisionato (VC-dimension, bias-variance trade-off)*
	  • *Comprensione specificità dei vari algoritmi di apprendimento*
	• *Autonomia di giudizio*:
	  • *Capacità di valutare la qualità della propria ricerca da un punto di vista statistico e della replicabilità dei risultati*
	• *Abilità comunicative:*
	  • *Capacità di comunicare i risultati scientifici della propria ricerca*
	• *Capacità di apprendere:*
	  • *Capacità di applicare gli strumenti illustrati a dati nuovi in completa autonomia.*

Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	*Il voto finale è attribuito in trentesimi. L'esame prevede una sola valutazione orale e si intende superato quando il voto è maggiore o uguale a 18.*

General information	
Year of the course	2023/2024
Academic calendar (starting and ending date)	*march 2024 –may 2024*
Credits (CFU/ETCS):	*6*
SSD	*Fis/07 o ING-INF/05*
Language	*Italian or English*
Mode of attendance	*Optional*

Professor/ Lecturer	
Name and Surname	Roberto Cilli
E-mail	*roberto.cilli@uniba.it*
Telephone	*327 730 8761*
Department and address	*Physics Department, Room 124*
Virtual room	*Italian or English*
Office Hours (and modalities: e.g., by appointment, on line, etc.)	*Every Friday, from 15 to 19 – by appointment*

Work schedule			
Hours			
Total	Lectures	Hands-on (laboratory, workshops, working groups, seminars, field trips)	Out-of-class study hours/ Self-study hours
150	*32*	*30*	*88*
CFU/ETCS			
6	*4*	*2*	

Learning Objectives	*The course will provide students with the necessary skills to understand and implement machine learning algorithms for knowledge extraction in large multi-omics data sets.*
Course prerequisites	*There are no specific prerequisites that differs from those required for admission to the master's degree course. However, prior knowledge of R-language is considered as an asset.*

Teaching strategies	*The teaching strategy follows the "learning-while-doing" philosophy with emphasis on the application of tools and learning theory to real-world case studies of interest to Bioinformaticians and Data Scientists. Therefore, the lessons consists in face-to-face lectures alternated coding classes.* *Classroom lectures rely on multimedia aids (PowerPoint presentations) and aim to teach the fundamentals of learning theory, supervised learning techniques, and data-driven model validation procedures.* *Coding classes are devoted to code R scripts performing all useful procedures for cleaning input data, performing preliminary analysis, and training and validating learning algorithms.*

Expected learning outcomes in terms of	
Knowledge and understanding on in terms of the Dublin descriptors	**Dublin Descriptor 1**: Knowledge and understanding skills; • Understanding of the general theorems of supervised learning theory and their implications. • Understanding of the specifics of machine learning algorithms introduced in the course. **Dublin Descriptor 2:** Ability to apply knowledge and understanding. • Ability to apply the analysis procedures explained in the course to conduct analyses on omics data never seen before. **Dublin Descriptor 3**: Critical and judgment skills. After completing, the student is expected to be able to: • apply the analysis procedures outlined in the course to conduct analyses on omics data never seen before. • make independent judgment about the classification performance measured for a given problem with reference to limitations or room for improvement of the proposed learning algorithm. • Understand ethical implications of various sources of bias in academic research. **Dublin Descriptor 4**: Ability to communicate what has been learned. After completing, the student is expected to have learnt to: • Communicate scientific results to their peers or stakeholders by using effective graphs. **Dublin Descriptor 5**: Ability to learn independently. After completing, the student is expected to have learnt to: • How to conduct empirical analyses on omics data never seen before by following the procedure outlined during the course
Syllabus	
Content knowledge	*• Introduction: what is Machine Learning; Problems, data, and tools.* *• An introduction to R.* *• Feature Engineering.* *• Dimensionality reduction; latent space methods; Principal Component Analysis (PCA);* *• Regression problem: linear regression, gradient descent.* *• Classification problems: Overfitting and complexity; bias and variance; training, validation, and test; Cross validation procedure; confusion matrix, accuracy, precision, recall, f1 score.* *• Linear classifiers; Naive Bayes model; Neural Networks, Support Vector Machine.* *• Decision Tree; Ensemble: Bagging, Random forests, Boruta, VSURF.* *• Unsupervised learning: clustering, k-means, hierarchical agglomeration.* *• Hints on Explainable Artificial Intelligence (XAI): SHAP Model.* *• Hints on CNNs and Unet-128.* *• Hints on Cohort Studies.*
Texts and readings	***Christopher M. Bishop, Pattern Recognition and Machine Learning***
Notes, additional materials	*PowerPoint presentations*
Repository	

Assessment	
Assessment methods	*Learning will be assessed by involving the students in a personal project consisting in conducting a data-driven analysis of omics data and by presenting the obtained results. The omics data should be adequately described by an international peer-reviewed article. More precisely, the project involves three steps:* *- software development and algorithm training* *- measurement and validation of classification/regression performance* *- creation of a 20-minute PowerPoint and oral interview (10 slides) regarding the performance of the proposed analysis chain* *The oral interview aims to understand whether the student has acquired the ability to independently apply the AI/ML tools and to defend their results. In particular, the assessment mode of the exam will pay special attention to measuring the student's ability to identify statistical association relationships in data of high dimensionality.*
Assessment criteria	*The student will be evaluated according to the following pillars.* *Knowledge and understanding skills:* • *Understanding of the general theorems of supervised learning theory;* *Applied knowledge and understanding skills:* • *Understanding implication of general theorems of supervised learning theory (VC-dimension, bias-variance trade-off)* • *Understanding specificity of various learning algorithms* *Autonomy of judgment:* • *Ability to assess the quality of one's research from a statistical perspective and replicability of results* *Communication skills:* • *- Ability to communicate scientific results to their peers* *Ability to learn:* • *- Ability to apply the tools explained to new data with complete autonomy.*
Final exam and grading criteria	*The examination involves only one oral assessment and is considered passed when the grade is greater than 18/30.*