COURSE OUTLINE

(1) GENERAL

SCHOOL	SCHOOL OF SCIENCES		
DEPARTMENT	DEPARTMENT OF CHEMISTRY		
LEVEL OF STUDIES	ISCED level 6 – Bachelor's or equivalent level		
COURSE CODE	EN17	**SEMESTER**	7th / 8th Semester
COURSE TITLE	Informatics II – Machine Learning		

TEACHING ACTIVITIES *If the ECTS Credits are distributed in distinct parts of the course e.g. lectures, labs etc. If the ECTS Credits are awarded to the whole course, then please indicate the teaching hours per week and the corresponding ECTS Credits.*	**TEACHING HOURS PER WEEK**	**ECTS CREDITS**
	3	3
Please, add lines if necessary. Teaching methods and organization of the course are described in section 4.		

COURSE TYPE *Background, General Knowledge, Scientific Area, Skill Development*	Skill Development
PREREQUISITES:	NO
TEACHING & EXAMINATION LANGUAGE:	GREEK
COURSE OFFERED TO ERASMUS STUDENTS:	NO
COURSE URL:	https://eclass2.emt.duth.gr

(2) LEARNING OUTCOMES

Learning Outcomes

Please describe the learning outcomes of the course: Knowledge, skills and abilities acquired after the successful completion of the course.

The aim of the course is to develop students' analytical thinking for extracting information from complex datasets and solving problems using data-driven methods and Machine Learning techniques. Upon successful completion of the course, students will be able to identify, select, and apply appropriate methods from multivariate data analysis and Machine Learning to problems requiring supervised and unsupervised learning. Regarding supervised learning, students will be able to apply methods to (i) regression problems using algorithms such as Least Squares Regression and Partial Least Squares Regression, and (ii) classification problems using algorithms such as Logistic Regression, Multinomial Logistic Regression, k-Nearest Neighbors, and Decision Trees. Regarding unsupervised learning, students will be able to apply methods to (i) dimensionality reduction problems and (ii) clustering problems using well-known algorithms from the literature. Students are expected to acquire practical skills in implementing these methodologies and techniques using the open-source programming language R, developing code within the RStudio Integrated Development Environment (IDE). The goal is to extract useful information hidden in raw data and evaluate the fitted models, thereby contributing to the assessment of claims and conclusions in data-driven solutions..

Upon successful completion of the course, students will be able to:

- Understand the fundamentals of Machine Learning, including supervised and unsupervised learning, basic terminology, and workflow for knowledge discovery in environmental chemistry, biology, and life and health sciences.

- Perform data preprocessing, including cleaning, transformation, normalization, and handling of missing values, and visualize univariate and multivariate data using R (histograms, density plots, boxplots, scatterplots, parallel coordinates, Chernoff faces, star plots, etc.).
- Apply supervised learning methods for regression problems, including simple and multiple linear regression, check model assumptions, perform diagnostics, fit models, and interpret results in R.
- Apply supervised learning methods for classification problems, including Logistic Regression, Multinomial and Ordinal Logistic Regression, Linear Discriminant Analysis, Naïve Bayes classifiers, Decision Trees, k-Nearest Neighbors, and Support Vector Machines, with practical case studies in R.
- Evaluate predictive performance using appropriate metrics for regression (errors, absolute error, relative error) and classification (accuracy, precision, recall, F-measure), and validate models using cross-validation (hold-out, k-fold, leave-one-out, leave-p-out), bootstrapping, and graphical methods such as regression error characteristic curves and ROC curves.
- Apply unsupervised learning techniques for clustering, including partitioning methods (k-means, k-medoids, EM) and hierarchical clustering (agglomerative and divisive algorithms) in R.
- Apply dimensionality reduction techniques such as Principal Component Analysis (PCA) and Correspondence Analysis to multivariate datasets.
- Combine multiple classifiers using ensemble methods, including ensemble averaging for regression, majority voting for classification, and bootstrap aggregating (bagging), with applications in R.
- Conduct experimental comparisons of algorithms, extract conclusions, and apply inferential statistics to assess model performance and generalization.
- Understand the principles of Explainable Artificial Intelligence (XAI), including the “black-box” problem, interpretability, explainability, transparency, and global/local explanation techniques, and apply these methods to real-world datasets in environmental chemistry, biology, and life and health sciences using R.

General Skills

Name the desirable general skills upon successful completion of the module

Search, analysis and synthesis of data and information,	*Project design and management*
ICT Use	*Equity and Inclusion*
Adaptation to new situations	*Respect for the natural environment*
Decision making	*Sustainability*
Autonomous work	*Demonstration of social, professional and moral responsibility and sensitivity to gender issues*
Teamwork	*Critical thinking*
Working in an international environment	*Promoting free, creative and inductive reasoning*
Working in an interdisciplinary environment	
Production of new research ideas	

Search, analysis and synthesis of data and information, Adaptation to new situations, Decision making, Autonomous work, Teamwork, Working in an international environment,

(3) COURSE CONTENT

Week 1: Introduction to Machine Learning, Classification Scheme of Machine Learning Techniques, Supervised Learning, Unsupervised Learning, Basic Terminology, Examples of Problems in Environmental Chemistry, Biology, Life and Health Sciences, Workflow for Knowledge Discovery.

Week 2: Data Preprocessing: Cleaning, Transformation, Normalization, Handling Missing Observations; Visualization of Univariate Data (Histogram, Probability Density Plot, q-q Plot, Boxplot, etc.); Visualization of Multivariate Data (Scatterplot, Parallel Coordinates Plot, Chernoff Faces, Star Plots, etc.). Case studies in Environmental Chemistry, Biology, Life and Health Sciences using the R programming language.

Week 3: Supervised Learning: Regression Problems, Simple and Multiple Linear Regression.

Week 4: Linear Regression Model Assumptions, Model Fitting, Model Interpretation, Diagnostic Checks. Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 5: Supervised Learning: Statistical Learning, Classification, Logistic Regression, Multinomial Logistic Regression, Ordinal Logistic Regression, Discriminant Analysis. Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 6: Supervised Learning: Classification, Naïve Bayes Classifier, Decision Trees for Classification and Regression, k-Nearest Neighbors, Support Vector Machines. Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 7: Predictive Performance Evaluation: Metrics for Regression Problems (Error, Absolute Error, Relative Error), Metrics for Classification Problems (Accuracy, Precision, Recall, F-measure); Validation Methods (Hold-out, k-fold, leave-p-out, leave-one-out cross-validation), Bootstrapping Simulation Methods, Graphical Validation Methods, Regression Error Characteristic Curve, ROC Curve, Area Under the ROC Curve.

Week 8: Unsupervised Learning: Clustering Methods, Partitioning Clustering, k-means, EM and k-medoids Algorithms, Hierarchical Clustering – Agglomerative and Divisive Algorithms. Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 9: Unsupervised Learning: Dimensionality Reduction Techniques, Principal Component Analysis, Correspondence Analysis.

Week 10: Combining Multiple Classifiers, Bias-Variance Tradeoff, Ensemble Averaging for Regression Problems, Voting Algorithm for Classification Problems, Bootstrap Aggregating (Bagging). Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 11: Experimental Comparison, Algorithm Comparison and Drawing Conclusions, Basic Concepts, Dataset Construction Process, Drawing Conclusions Using Inferential Statistics.

Week 12: Introduction to Explainable Artificial Intelligence (XAI), "Black Box" Problem, Definitions and Terminology, Comprehensibility, Understandability, Interpretability, Explainability, Transparency, Classification Scheme of XAI Methods, Global and Local Explainability Techniques. Case studies in Environmental Chemistry, Biology, Life and Health Sciences using R.

Week 13: Presentation of the Machine Learning Life-cycle in real-life datasets.

(4) LEARNING & TEACHING METHODS - EVALUATION

TEACHING METHOD *Face to face, Distance learning, etc.*	Face to face
USE OF INFORMATION & COMMUNICATIONS TECHNOLOGY (ICT) *Use of ICT in Teaching, in Laboratory Education, in Communication with students*	Use of ICT in Teaching Use of ICT in Communication with students
TEACHING ORGANIZATION *The ways and methods of teaching are described in detail.* *Lectures, Seminars, Laboratory Exercise, Field Exercise, Bibliographic research & analysis, Tutoring, Internship (Placement), Clinical Exercise, Art Workshop, Interactive learning, Study visits, Study / creation, project, creation, project. Etc.* *The supervised and unsupervised workload per activity is indicated here, so that total workload per semester complies to ECTS standards.*	(see table below)

Activity	Workload/semester
Lectures	39
Study	33
Exams	3
Total	75

STUDENT EVALUATION *Description of the evaluation process* *Assessment Language, Assessment Methods, Formative or Concluding, Multiple Choice Test, Short Answer Questions, Essay Development Questions, Problem Solving, Written Assignment, Essay / Report, Oral Exam, Presentation in audience, Laboratory Report, Clinical examination of a patient, Artistic interpretation, Other/Others* *Please indicate all relevant information about the course assessment and how students are informed*	Written final exam that includes problem solving from different sections of the course.

(5) SUGGESTED BIBLIOGRAPHY

– Στατιστική και Μηχανική Μάθηση με την R. Ιωαννίδης Δ., Αθανασιάδης Ι. Εκδόσεις ΤΖΙΟΛΑ. ISBN: 978-960-418-642-6
– An Introduction to Statistical Learning with Applications in R. (Second Edition) G. James, D. Witten, T. Hastie, R. Tibshirani. Springer