

COURSE DESCRIPTION – ACADEMIC YEAR 2017/2018

Course title	Advanced Statistics for Data Mining
Course code	76001
Scientific sector	MAT/06
Degree	European Master in Software Engineering (LM-18)
Semester	1
Year	1
Credits	12
Modular	Yes
University	UniBZ

Total lecturing hours	72
Total lab hours	24
Total exercise hours	12
Attendance	Not compulsory
Prerequisites	Students should have a good knowledge of probability and algorithms and suitable programming skills for projects.
Course page	https://ole.unibz.it/

Specific educational objectives	<p>The course belongs to the type "affine" and is part of the Transversal Skills offered from unibz (EMSE-TS).</p> <p>It is designed to give a general overview of scientific contents about Data Mining.</p> <p>This course aims at developing a deep understanding of the strengths and the limitations of a wide range of data mining techniques to be able to identify their use cases and important applications. Given a data-mining problem, students will be able to define what are the data sources to be exploited, the mining tasks to be performed, and the algorithms that need to be used to solve the problem. To this end, students will have the possibility to actively participate in data mining projects to perform extensive experiments on real datasets and potentially propose extensions to existing algorithms. A useful takeaway from the course will be the ability to identify the knowledge to be mined from data and employ mining algorithms using easy-to-use software and cases.</p>
--	--

Module 1	Data Mining
Module code	76001A
Module scientific sector	INF/01
Lecturer	Mouna Kacimi El Hassani
Contact	Piazza Domenicani 3, Room 2.02 , Mouna.Kacimi@unibz.it , +39 0471 016114
Scientific sector of lecturer	INF/01
Teaching language	English
Office hours	To be arrange beforehand by email or during the lecture.
Lecturing assistant (if any)	--
Office hours LA	--
Credits	8
Lecturing hours	48
Lab hours	24

Exercise hours	--
List of topics	<ul style="list-style-type: none"> • Data types, quality and pre-processing • Data exploration • Classification • Association analysis • Clustering • Error estimation • Data mining applications • Data mining tools
Teaching format	Frontal lectures and labs with theoretical and practical exercises. Group projects.

Module 2	Statistical Methods
Module code	76001B
Module scientific sector	MAT/06
Lecturer	Leonardo Ricci
Contact	Piazza Domenicani 3, Room 1.04, leonardo.ricci@unibz.it
Scientific sector of lecturer	FIS/01
Teaching language	English
Office hours	During the lecture time span
Lecturing assistant (if any)	--
Office hours LA	--
Credits	4
Lecturing hours	24
Lab hours	--
Exercise hours	12
List of topics	<ul style="list-style-type: none"> • Discrete random variables and their distributions • Statistical Inference • Correlation and regression • Time series analysis • Dynamic systems and Markov chains
Teaching format	Frontal lectures and project work during the exercise hours.

Learning outcomes	<p>Knowledge and understanding:</p> <ul style="list-style-type: none"> • Know in detail the scientific methodology of investigation applied to information systems and to innovative technologies with a particular focus on empirical data analysis. Understand and use tools of discrete and continuous mathematics, applied mathematics and physics, which are of support to informatics and its mathematical applications. <p>Applying knowledge and understanding:</p> <ul style="list-style-type: none"> • Be able to apply empirical analysis fundamentals of ICT data (e.g. Data Mining) for the construction of mathematical models for the evaluation and prediction of the applications' features and software systems • Be able to plan and conduct information systems' analyses in order to acquire measures related to their behaviour and evaluate experimental hypothesis in the industrial sector or in the applied research
-------------------	---

Making judgments

- Be able to autonomously select documents from different sources, including technical books, digital libraries, scientific technical journals, web portals, open source software and hardware
- Be able to identify work goals, compatible with available time and resources.

Communication skills

- Be able to present in a fixed time the content of a scientific / technical report to an audience composed also of non-specialists
- Be able to coordinate a project team and identify those activities needed to reach the goals of the project

Learning skills

- Be able to autonomously extend the knowledge acquired during the study course by reading and understanding scientific and technical documentation in English
- Be able to formulate and validate theories and define new methods through empirical induction

Module 1: Data Mining

Knowledge and understanding

- Know the main techniques for extracting information (associations, trends, dependencies, forecasts) from structured and unstructured data.
- Understand the methods of mathematics and statistics which are of support to information technology and its applications.

Applying knowledge and understanding

- Be able to identify new application requirements and business opportunities in the field of systems based on data and knowledge.
- Be able to design and execute experimental analyses on information systems or their components.

Making judgments

- Be able to independently select the documentation required to keep abreast of the frequent technological innovations in the field by using a wide variety of documentary sources: books, web, magazines.
- Be able to identify reasonable work goals and estimate the resources required to achieve the objectives.

Communication skills

- Be able to structure and prepare scientific and technical documentation describing project activities.
- Be able to coordinate the work of a project team and to interact positively with members of the group.

Ability to Learn

- Be able to autonomously extend the knowledge acquired during the study course by reading and understanding scientific and technical documentation.
- Be able to independently keep up to date with developments in the most important areas of Data Mining.

	<p>Module 2: Statistical Methods</p> <p>Knowledge and understanding:</p> <ul style="list-style-type: none"> • Thoroughly understand the scientific method of investigation. • Understand methods of mathematics and of statistics that support Information Technology and its applications. <p>Applying knowledge and understanding:</p> <ul style="list-style-type: none"> • Be able to design and execute experimental analyses on information systems or their components. <p>Making judgments:</p> <ul style="list-style-type: none"> • Be able to work autonomously according to the own level of knowledge <p>Communication skills:</p> <ul style="list-style-type: none"> • Be able to structure and write scientific documentation. <p>Learning skills:</p> <ul style="list-style-type: none"> • Have developed learning capabilities to pursue further studies with a high degree of autonomy. • Be able to learn the innovative features of state-of-the-art technologies and information systems
Assessment	<p>Module 1: Data Mining</p> <ul style="list-style-type: none"> • Final written exam: with verification questions and problem solving tests • Project: consists in two parts (1) acquiring the basic data mining knowledge via assignments and (2) applying or adapting existing techniques to solve a concrete problem chosen by the student. The project results are a written report, implementation of the used techniques, and a presentation. • Mid-term exam (optional): with verification questions and problem solving tests <p>Module 2: Statistical Methods</p> <p>Written final exam only [100 % of mark]. The exam consists of 4-6 exercises.</p>
Assessment language	English
Evaluation criteria and criteria for awarding marks	<p>Module 1: Data Mining</p> <p>Evaluation Criteria</p> <ul style="list-style-type: none"> • Written final exam: 50% of the mark • Project: 50% of the mark • The project marks are valid for the three exam sessions. The project is a prerequisite for attending the written exam <p>Criteria for awarding marks</p> <ul style="list-style-type: none"> • Exam: correctness and clarity of answers, the ability to adequately solve data mining problems and to understand how to choose the right technique. • Project: ability to apply data mining tools and algorithm in a real world problem, creativity, and ability to work in team.

	<p>Module 2: Statistical Methods</p> <p>Correctness of answers / calculations</p> <p>The final mark given for the course results from the weighted arithmetical mean of the marks obtained in each of the two examinations, following the formula $(2/3) * \text{DataMining_mark} + (1/2) * \text{AdvancedStatistics_mark}$</p>
Required readings	<p>Module 1: Data Mining</p> <p>Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, 2006</p> <p>Module 2: Statistical Methods</p> <ol style="list-style-type: none"> 1. D. Freedman, R. Pisani, R. Purves, "Statistics" (International student edition, 4th edition), W.W. Norton & Company, 2007; 2. I. Miller, M. Miller, "John E. Freund's Mathematical Statistics with Applications" (7th Edition), Pearson; 3. M. Baron, "Probability and Statistics for Computer Scientists" (1st edition), Chapman and Hall/CRC, 2006.
Supplementary readings	<p>Additional sources will be announced during the course.</p>
Software used	<ul style="list-style-type: none"> • R/RStudio, Python+ScikitLearn/Jupyter