DATA MINING

SYLLABUS
2021-2022

INSTRUCTOR INFORMATION	FERNANDO LUCAS BAÇÃO 2º floor, room 10 Tel: 21 3870413 (ext. 222) bacao@novaims.unl.pt http://www. novaims.unl.pt/fbacao JOÃO FONSECA jpfonseca@novaims.unl.pt DAVID SILVA dsilva@novaims.unl.pt
SCHEDULE	MAA-DS Theoretical Sessions • Thursday from 15h30 – 16h45 Practical Sessions (João Fonseca) • P5 - Tuesday from 14h00 – 15h30 • P2 – Wednesday from 09h30 – 11h00 • P4 - Thursday from 11h00 – 12h30 MAA-BA Theoretical Sessions • Wednesday from 13h00 – 14h15 Practical Sessions (David Silva) • P1 - Wednesday from 11h00 – 12h30 • P6 - Wednesday from 13h00 – 14h30 • P3 - Wednesday from 16h00 – 17h30
OFFICE HOURS:	Wednesday from 12h00 – 13h00 (schedule appointment by email), 2nd Floor, Room 10
CONTACT	The course has its own email address mdsaa2021@gmail.com, which should be used by the student to contact the teachers. To submit any homework and/or projects you must use Moodle.
DESCRIPTION	The Data Mining course aims to study the main methods and tools available in data mining (knowledge discovery in databases), more specifically the subset of tools which are usually called descriptive models (or unsupervised learning). The course does not assume familiarity of the student with the theme, but it is highly recommended that

	the students have knowledge of inferential statistics, as well as minimal computer skills. The course seeks to achieve a balance between courses dedicated to in-depth analysis of the algorithms and the courses for managers that seek to raise awareness about the importance of the tools. This is a technical course for all, those who already work or want to work in developing descriptive models and exploring big databases. As such, students will perform the activities of a typical data scientist, especially in the project work, which constitutes a central component of the course. The main focus in this course is to present the algorithms in a clear and comprehensible way to a wide audience with different academic backgrounds. It is intended to enable the student to understand the fundamentals associated with the inner workings of the different algorithms, because only then the student will be able to apply them judiciously. The course program covers the main methodological aspects, data preparation and preprocessing tasks as well as the most popular descriptive models, including different clustering algorithms and association rules, among others. The aim is also to provide students the opportunity to learn/use Python to implement and apply these algorithms in real world applications.
OBJECTIVES	At the end of the course, students should be able to: • Discuss the most relevant ideas and concepts associated with data mining; • Understand the fundamentals of exploratory data analysis, including the use of graphics, both to present and analyze data; • Be able to execute basic and intermediate data preparation and pre-processing tasks (e.g. detect outliers or dealing with missing values); • Describe and use Multidimensional Visualization Methods, such as such as principal components analysis, t-SNE, Umap and self-organizing maps; • Describe with detail segmentation techniques such as cohort analysis and RFM analysis; • Describe with detail clustering techniques such as hierarchical clustering, partitioning methods (k-means and meadoids), density-based methods and fuzzy clustering;

	• Understand the trade-offs involved in the definition of the number of clusters and how to interpret and analyze a clustering solution; • Discuss the use of nearest neighbors and decision trees to explore and get insights on clustering solutions; • Create a segmentation, being able to explain the options used and explaining alternative approaches, whenever available; • Describe the *apriori* algorithm, as well as calculate and explain the most relevant performance measures of association rules;
COURSE SUCCESS	In this course success depends on a number of factors: • Basic knowledge of statistics; • Attend classes; • Work during the semester and not only when the exams are about to start; • Develop the course project during the semester, making the most of the practical classes; • Read the suggested references.
CONTENTS	1. Introduction to the Data Mining Course a. Syllabus b. Objectives c. Course projects d. Grading e. Bibliography 2. The Context a. The Growth of the Digital Universe b. Buzz Words and Definitions c. Data Science d. Different Roles in Data Science e. The relevance of Data 3. The canonical tasks in Data Mining and work process a. Canonical tasks in Data Mining i. Supervised Learning ii. Unsupervised learning b. The Data Mining Process i. KDD process ii. The CRISP DM Methodology iii. The SEMMA Methodology c. Before starting analysis i. Types of Measurements ii. Problem definition

4. Exploratory Data Analysis
    a. Univariate
        i. Categorical
        ii. Numerical
    b. Bivariate
        i. Categorical
        ii. Numerical and Numerical
        iii. Categorical and Numerical
    c. Graphics
        i. Information Visualization Guidelines
        ii. Graphics for Presentation
        iii. Graphics for Analysis
5. Data Preparation and Preprocessing
    a. Data Preparation
        i. Noisy Data
        ii. Missing Values
        iii. Outlier Detection
        iv. Data discretization
        v. Imbalanced learning
    b. Data Preprocessing
        i. The curse of dimensionality
        ii. Dimensionality reduction principles
        iii. Input Space Reduction – Relevancy
        iv. Input Space Reduction – Redundancy
        v. Data Standardization
6. Data Segmentation Strategies
    a. Cohort analysis
    b. Cell-based segments
        i. two-way
        ii. over time
    c. RFM analysis
7. Data Clustering
    a. Motivation
    b. Definition and Notations
    c. Similarity Measurements
    d. Clustering Techniques
        i. Hierarchical algorithm
        ii. K-means algorithm
        iii. Nearest Neighbor Clustering
        iv. Density-based
        v. Fuzzy
        vi. Evolutionary
    e. Analysis and validation of clustering solutions

	i. The number of clusters ii. Analysis and profiling of the clustering solution f. Clustering Strategies i. Hiearchical – partition ii. Partition – hierarchical g. Semi-Supervised Classification i. Classification trees ii. K-nearest neighbour 8. Multidimensional Visualization Methods a. Principal Component Analysis b. tSNE c. Umap d. Self-Organizing Maps 9. Association Rules a. Motivation (market basket analysis) b. Frequent Itemsets c. Association Rules Measures i. Support ii. Confidence iii. Lift d. Association Rules Algorithms i. Apriori Algorithm ii. Improving the Efficiency of Apriori e. From Association Mining to Correlation Analysis
BIBLIOGRAPHY	References: ❑ Han, J., Kamber, M. 2006, Data Mining – Concepts and Techniques, Morgan Kaufmann, Elsevier Inc. ❑ A. K. Jain, M.N. Murthy and P.J. Flynn, 1999 Data Clustering: A Review, ACM Computing Review. Additional References: ❑ Mitchell, T., (1997) Machine Learning, McGraw Hill. ❑ Provost, F., Fawcett, T. (2013) Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media, ISBN-13 : 978-1449361327. ❑ Berry, M.J.A. Linoff, G., 1997, Data Mining Techniques for marketing, sales and customer support. Second Edition. 2004, John Wiley & Sons. ❑ Bishop (2006) Pattern Recognition and Machine Learning, Springer, ISBN-13 : 978-0387310732.

	Note: all references are available at NOVA IMS library or are provided by the teacher.		
EVALUATION	1ª Session – Exam (65%), Project (35%) 2ª Session – Exam (65%), Project (35%)		
CALENDAR	L 1	8&9 Sept.	Introduction to the Data Mining Course Syllabus Objectives Course projects Grading Bibliography The Context The Growth of the Digital Universe Buzz Words and Definitions Data Science Different Roles in Data Science The relevance of Data
	No practical sessions in the 1st week		
	L 2	15&16 Sep.	The canonical tasks in Data Mining and work process Canonical tasks in Data Mining Supervised Learning Unsupervised learning The Data Mining Process KDD process The CRISP DM Methodology The SEMMA Methodology Before starting analysis Types of Measurements Problem definition
	Practical sessions (Python)		
	L 3	22&23 Sep.	Exploratory Data Analysis Univariate Categorical Numerical Bivariate Categorical Numerical and Numerical Categorical and Numerical
	Practical sessions (Python)		
	L 4	29&30 Sep.	Exploratory Data Analysis Graphics Information Visualization Guidelines Graphics for Presentation Graphics for Analysis
	Practical sessions (Python)		
	L 5	6&7 Oct.	Data Preparation and Preprocessing Data Preparation Noisy Data Missing Values

		Outlier Detection Data discretization Imbalanced learning
Practical sessions (Python)		
L 6	13&14 Oct.	Data Preparation and Preprocessing Data Preprocessing The curse of dimensionality Dimensionality reduction principles Input Space Reduction – Relevancy Input Space Reduction – Redundancy Data Standardization
Practical sessions (Python)		
L 7	20&21 Oct.	Data Segmentation Strategies Cohort analysis Cell-based segments two-way over time RFM analysis
Practical sessions (Python)		
L 8	3&4 Nov.	Data Clustering Motivation Definition and Notations Similarity Measurements Clustering Techniques Hierarchical Partitional
Practical sessions (Python)		
L 9	10&11 Nov.	Data Clustering Clustering Techniques Density-based Fuzzy clustering Neural Networks Evolutionary
Practical sessions (Python)		
L 10	17&18 Nov.	Data Clustering Analysis and validation of clustering solutions The number of clusters Analysis and profiling of the clustering solution Clustering Strategies Hierarchical – partition Partition – hierarchical Semi-Supervised Classification Classification trees K-nearest neighbour
Practical sessions (Python)		

L 11	24&25 Nov.	Multidimensional Visualization Methods Principal Component Analysis tSNE algorithm
Practical sessions (Python)		
L 12	**4**&2 Dec.	Multidimensional Visualization Methods UMap algorithm Self-Organizing Maps
Practical sessions (Python)		
L 13	11&9 Dec.	Association Rules Motivation (market basket analysis) Frequent Itemsets Association Rules Measures Support Confidence Lift Association Rules Algorithms Apriori Algorithm Improving the Efficiency of Apriori From Association Mining to Correlation Analysis
Practical sessions (Python)		
L 14	15&16 Dec.	Course Overview Exam preparation

Course Projects

Project consists of a practical clustering application using Python. In this project the students will complete the segmentation of a customers database, following all the usual steps of a real world project. For this the students will receive a set of specific guidelines that they should follow. The guidelines provide information about the type of tasks the students should do and the general results they should achieve. The end product of the project should be a report about the database and the different customer segments of the company. With this project the students should develop their analytical skills, but also their proficiency in working with large datasets, extracting, transforming and loading tasks and visualization and reporting.

Project discussion: after submitting the projects the students will be called to discuss the project with one of the instructors.

Project groups: the project can be done individually or in groups (the latter is a better option) the groups should not exceed 3 students.

Project Deadline: January 4th

Tasks. In both, practical and theoretical classes, students will be frequently assigned homework, which will consist of simple tasks related with the course material. It is expected that the students complete these tasks.

Final Exam. The exam will be a single hour in-class exam covering all the material of the course. The exam will consist of 15to 20 multiple-choice questions, 5 to 10 true or false questions and a small essay.

Grading

Project: 35%
Exam: 65%

Both components of the evaluation (project and exam) are mandatory. There are two opportunities to do the exam. Any delay in the delivery of the project is subject to a penalty of 10% of the grade for each day of delay. Please note that the project will be developed in groups, but each group cannot have more than 3 elements. To obtain approval in the discipline the student **cannot have less than 8 (40%) in the exam grade**.