DATA MINING

SYLLABUS
2023-2024

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 FARINA PONTEJOS fpontejos@novaims.unl.pt
SCHEDULE	TP1 Theoretical Sessions • Tuesdays from 14h00 – 15h15 Practical Sessions (Farina Pontejos) • P1 - Tuesday 15h30 - 17h00 • P2 - Wednesday 15h30 - 17h00 • P3 - Wednesday 14h00 - 15h30 TP2 Theoretical Sessions • Tuesdays from 15h30 – 16h45 Practical Sessions (João Fonseca) • P4 - Wednesday 10h00 - 11h30 • P5 - Wednesday 15h30 - 17h00 • P6 - Tuesday 14h00 - 15h30
OFFICE HOURS:	Tuesday from 13h00 – 14h00 (schedule appointment by email), 2nd Floor, Room 10
CONTACT	All communications with the instructors should be done using the Moodle platform. To submit any homework and/or projects you must also use Moodle.
DESCRIPTION	The goal of the Data Mining course is to introduce students to the primary techniques and tools used in data mining, particularly those categorized as descriptive models or unsupervised learning. While no prior familiarity with the subject is assumed, it is strongly recommended that students have a basic understanding of inferential statistics and some minimal computer skills. This course strives to strike a balance between offering in-depth analysis of algorithms and providing managerial insights into the significance of these tools. It caters to a wide audience, encompassing individuals already engaged in or interested in pursuing roles related to creating descriptive models and exploring large databases. Consequently, students will engage in activities typical of a data scientist, with a particular emphasis on the central role of project work.

	The primary focus of this course is to explain the algorithms in a manner that is clear and comprehensible to a diverse academic audience. The intention is to equip students with a fundamental understanding of how these algorithms function internally, as only then can they be applied judiciously. The course curriculum encompasses key methodological aspects, data preparation, and preprocessing tasks, along with the most popular descriptive models such as various clustering algorithms and association rules, among others. Additionally, the course aims to provide students with the opportunity to learn and utilize Python for implementing and applying these algorithms in real-world scenarios.
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 medoids), and fuzzy clustering; • Describe with detail density-based clustering techniques such as DBSCAN and Mean-Shift; • 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. Introduction to Data Science
   a. Data as a strategic resource
   b. Definitions
      i. Artificial intelligence
      ii. Machine learning
      iii. Big data
      iv. Data Science
   c. Data Science roles and skills
      i. Data scientist
      ii. Data engineer
      iii. Data analyst
   d. Fundamental principles of data-driven thinking
      i. The process of developing a model
      ii. The role of features
      iii. The importance of data
      iv. Statistics versus Data Science
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 algorithms
        - ii. Partitional algorithms (k-means and k-medoids)
        - iii. Iterative Self-Organizing Data Analysis Technique (ISODATA)
        - iv. Density-based algorithms (DBSCAN and Mean-Shift)
        - v. Mean Shift algorithm
        - vi. Fuzzy clustering
        - vii. Evolutionary algorithms for clustering
    - e. Analysis and validation of clustering solutions
        - i. The number of clusters
        - ii. Analysis and profiling of the clustering solution
    - f. Clustering Strategies
        - i. Hieararchical – partition
        - ii. Partition – hierarchical
    - g. Semi-Supervised Classification
        - i. Classification trees
        - ii. K-nearest neighbour
8. Multidimensional Visualization Methods
    - a. Principal Component Analysis
    - b. t-Distributed Stochastic Neighbor Embedding (t-SNE)
    - c. Uniform Manifold Approximation and Projection (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.

<table>
<tr><td></td><td colspan="3">□ A. K. Jain, M.N. Murthy and P.J. Flynn, 1999 Data Clustering: A Review, ACM Computing Review.
□ 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.

Additional References:
□ Mitchell, T., (1997) Machine Learning, McGraw Hill.
□ 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.</td></tr>
<tr><td>EVALUATION</td><td colspan="3">1ª Session – Exam (65%), Project (35%)
2ª Session – Exam (65%), Project (35%)</td></tr>
<tr><td>CALENDAR</td><td>L 1</td><td>5 Sept.</td><td>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</td></tr>
<tr><td></td><td colspan="3">Practical sessions (Python)</td></tr>
<tr><td></td><td>L 2</td><td>12 Sep.</td><td>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</td></tr>
<tr><td></td><td colspan="3">Practical sessions (Python)</td></tr>
<tr><td></td><td>L 3</td><td>19 Sep.</td><td>Exploratory Data Analysis
Univariate
Categorical
Numerical
Bivariate
Categorical</td></tr>
</table>

		Numerical and Numerical Categorical and Numerical
Practical sessions (Python)		
L 4	26 Sep. No Class	Exploratory Data Analysis Graphics Information Visualization Guidelines Graphics for Presentation Graphics for Analysis
Practical sessions (Python)		
L 5	3 Oct.	Data Preparation and Preprocessing Data Preparation Noisy Data Missing Values Outlier Detection Data discretization Imbalanced learning
Practical sessions (Python)		
L 6	10 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	17 Oct.	Data Segmentation Strategies Cohort analysis Cell-based segments two-way over time RFM analysis
Practical sessions (Python)		
L 8	31 Oct.	Data Clustering Motivation Definition and Notations Similarity Measurements Clustering Techniques Hierarchical Partitional
Practical sessions (Python)		
L 9	7 Nov.	Data Clustering Clustering Techniques Density-based Mean Shift algorithm Fuzzy clustering Evolutionary
Practical sessions (Python)		
L 10	14 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	21 Nov.	Multidimensional Visualization Methods Principal Component Analysis t-SNE algorithm
Practical sessions (Python)		
L 12	28 Nov.	Multidimensional Visualization Methods UMap algorithm Self-Organizing Maps
Practical sessions (Python)		
L 13	5 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	12 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 customer 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 7th

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 15 to 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**.