

Ficha de Unidade Curricular – (Versão A3ES 2018-2023)

1. Caracterização da Unidade Curricular.

- 1.1. **Designação da unidade curricular** (1.000 carateres).
Engenharia nos Megadados / Big Data Engineering
- 1.2. **Sigla da área científica em que se insere** (100 carateres).
IC
- 1.3. **Duração**¹ (100 carateres).
Semestral
- 1.4. **Horas de trabalho**² (100 carateres).
162
- 1.5. **Horas de contacto**³ (100 carateres).
67.5H (T: 22.5H TP: 24H ; PL: 21H)
- 1.6. **ECTS** (100 carateres).
6
- 1.7. **Observações**⁴ (1.000 carateres).
UC optativa, MEIC
- 1.7. **Remarks** (1.000 characters).
Elective, MEIC

2. Docente responsável e respetiva carga letiva na Unidade Curricular (preencher o nome completo) (1.000 carateres). Matilde Pós-de-Mina Pato (67.5)

3. Outros docentes e respetivas cargas letivas na unidade curricular (1.000 carateres).

4. Objetivos de aprendizagem (conhecimentos, aptidões e competências a desenvolver pelos estudantes). (1.000 characters).

Os estudantes que terminam com sucesso esta unidade curricular serão capazes de:

1. Compreender as técnicas de modelação de dados multidimensionais.
2. Conhecer os diferentes sistemas para armazenamento e processamento de dados em larga escala no suporte a sistemas de apoio à decisão (SAD).
3. Selecionar as metodologias apropriadas para os processos de extracção, transformação e carregamento e aplicá-las na resolução de problemas reais de análise de dados e tomada de decisão.
4. Conhecer as diferentes arquiteturas de armazenamento para megadados.
5. Identificar as diferentes fontes de dados, os tipos de dados, a qualidade dos dados e as características dos e as transformações necessárias em determinado contexto analítico;
6. Compreender a importância dos metadados (catálogo de dados) no contexto do processamento analítico para a tomada de decisão.
7. Analisar os resultados de ferramentas de Processamento Analítico e monitorização de indicadores, relatórios e *dashboarding*.
8. Escrever relatórios técnicos e elaborar apresentações técnicas com análise comparativa e discussão de diferentes resultados.

4. Intended learning outcomes (knowledge, skills, and competencies to be developed by the students). (1.000 characters).

Students who complete this course unit will be able to:

1. Understand multidimensional data modelling techniques.
2. Know the different systems for storing and processing large-scale data in support of decision support systems (DSS).

3. Select the appropriate methodologies for the extraction, transformation and loading processes: ELT and ETL, and apply them to solving real problems of data analysis and decision-making.
4. Know the different storage architectures for mega data: Data Warehouse, Data Lake and Data Mesh.
5. Identify the different data sources, data types, data quality and characteristics (data quality and data profiling) and the necessary transformations in a given analytical context.
6. Understand the importance of metadata (data catalogue) in the context of analytical processing for decision-making.
7. Analyse the results of analytical processing tools and monitor indicators, reports and dashboards.
8. Write technical reports and prepare presentations with comparative analysis and discussion of different solutions.

5. Conteúdos programáticos (1.000 carateres).

- I. Introdução aos Sistemas de Suporte à Decisão (SSD)
- II. Objetivos, diferentes tipos e desafios dos SSD. Importância dos dados nos processos de decisão.
- III. Análise de dados e modelação multidimensionais.
- IV. Análise de dados com ferramentas OLAP. Objetivos da modelação dimensional. Diferentes tipo de modelos multidimensionais. Metodologias de passagem de modelos entidade-associação para modelos multidimensionais: OLAP, ROLAP e MOLAP.
- V. Processos de extração, transformação e carregamento de dados (ELT e ETL). Importância destes processo e diferentes abordagens. *Workflow* típico no processamento de dados. Integração de dados. Ferramentas de modelação.
- VI. Arquiteturas de dados para *Big Data*.
- VII. Limitações para armazenamento de dados dinâmicos e pouco estruturados. Arquiteturas *Data Lake* e *Data Mesh*. Coexistência e mapeamento entre diferentes arquiteturas de armazenamento.
- VIII. Catalogação de dados e preparação destes para exploração. Integração, limpeza (*data cleaning* ou *data scrubbing*), normalização e transformação de dados. Tratamento de dados duplicados e em falta (*Data wrangling*). *Data augmentation*.

5. Syllabus (1.000 characters).

- I. Introduction to Decision Support Systems (DSS).
- II. Objectives of a DSS. Different types of DSS. Importance of data in decision-making. Challenges facing the DSS.
- III. Multidimensional data analysis and modelling
- IV. Data analysis with OLAP tools. Objectives of dimensional modelling. Different types of multidimensional models. Methodologies for moving from entity-association models to multidimensional models: OLAP, ROLAP and MOLAP.
- V. Processes of data extraction, transformation and loading (ELT and ETL). The importance of these processes and methods. Typical workflows in the data processing. Data integration. Modelling tools.
- VI. Data architectures for Big Data.
- VII. Limitations of data warehouses for storing dynamic and unstructured data. Data Lake and Data Mesh architectures. Coexistence and mapping between different storage architectures.
- VIII Cataloguing data and preparing it for exploitation. Data integration, data cleaning & data scrubbing, normalisation and transformation. Handling duplicate and missing data (Data wrangling). Data augmentation.

6. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular (1.000 carateres).

Esta UC constitui o primeiro contacto dos estudantes com a área de engenharia de dados para suporte à tomada de decisão, dotando os estudantes de conhecimentos fundamentais que lhes permitam perceber a importância destes e o modo como as tecnologias da informação suportam a recolha, o tratamento e a análise dos dados em contextos organizacionais de SAD. Os tópicos (I) e (II) introduzem a cultura geral necessária para compreender o tema; os tópicos (III) a (VIII) permitem aferir o cumprimento dos objectivos de aprendizagem (3) a (7). Com a realização e apresentação do trabalho prático e a elaboração do respetivo relatório é possível aferir o objetivo de aprendizagem (8).

6. Evidence of the syllabus's coherence with the curricular unit's intended learning outcomes (1.000 characters).

The course introduces students to the area of data engineering for decision support. It provides students with the fundamental knowledge that will enable them to understand the importance of data and how information technologies support the collection, processing and analysis of data in the organisational contexts of DSS. Topics (I) and (II) introduce the general culture required to understand the subject; cases (III) to (VIII) enable the assessment of the achievement of learning objectives (3) to (7). As part of understanding objective (8), practical work and the corresponding report must be completed and presented.

7. Metodologias de ensino (avaliação incluída) (1.000 carateres).

Metodologia de ensino é baseada na abordagem *Problem-Based Learning* (PBL). Pretende-se privilegiar a autonomia do estudante no desenvolvimento de soluções para problemas complexos, adequados ao seu nível cognitivo. Incentiva-se o trabalho em grupo e a discussão/reflexão em sessões de grupo. As aulas destinam-se à apresentação dos temas e de exemplos práticos de aplicação. Os objetivos de aprendizagem de (1) a (7) são avaliados através da componente teórica (CT), constituída por avaliação presencial. Os objetivos de aprendizagem (1) a (8) são avaliados através da componente prática (CP), que consiste na realização de um trabalho prático (TP), pedagogicamente fundamental, a escrita do relatório e uma apresentação em contexto de sala. A classificação final é obtida através da média aritmética simples de ambas as componentes. Para aprovação na UC, a classificação mínima da CT é de 8 valores. A avaliação em épocas especiais consiste na elaboração de 2 TPs e de 1 exame escrito, valendo cada uma destas componentes 50% da nota final.

7. Teaching methodologies (including assessment) (1.000 characters).

The teaching methodology is based on the Problem-Based Learning (PBL) approach. It is intended to privilege student autonomy in developing solutions to complex problems appropriate to their cognitive level. Group work and discussion/reflection are encouraged in group sessions. Classes are designed to present the topics and practical examples of application. Learning objectives (1) to (7) are assessed using the theoretical component (CT), consisting of a face-to-face evaluation. Learning objectives (1) to (8) are evaluated through the practical part (CP), which consists of pedagogically actual practical work (TP), the writing of a report and a presentation in class. The final classification is obtained through the simple arithmetic mean of both components. For approval in the UC, the minimum mark of the CT is 8 points. Special-season evaluation consists of 2 TPs and one written exam, each component worth 50% of the final mark.

8. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular (3.000 carateres).

As aulas destinam-se à apresentação das bases teóricas dos conteúdos programáticos (aulas teóricas). Nas aulas, são desenvolvidos pequenos projetos e analisados casos de estudo (aulas teórico-práticas). Privilegia-se uma forma de apresentação interativa. A componente laboratorial (aulas práticas) serve para aplicar num ambiente controlado as técnicas apresentadas. O trabalho autónomo (extra-aula) é guiado pelo trabalho prático (projeto), concebido para consolidar as competências de conceção e desenvolvimento dos conteúdos programáticos. O projeto é apresentado aos estudantes no início do semestre guiando os exemplos e tópicos lecionados. Os objetivos de aprendizagem são identificados nos guiões apresentados aos estudantes, permitindo clarificar as competências que são necessárias adquirir no desenvolvimento do projeto e nas aulas práticas.

8. Evidence of the teaching methodologies' coherence with the curricular unit's intended learning outcomes (3.000 characters).

The classes aim to present the theoretical basis of the course contents (academic classes). In class, small projects are developed, and case studies are analysed (theoretical-practical classes). An interactive form of presentation is favoured. The laboratory component (practical courses) serves to apply the techniques presented in a controlled environment. Autonomous work (extra-class) is guided by experimental work (project) designed to consolidate the design and development skills of the course contents. The project is presented to students at the beginning of the semester, guiding the examples and topics taught. The learning objectives are identified in the guides given to the students, allowing clarification of the skills necessary to acquire in the development of the project and the practical classes.

9. Bibliografia de consulta/existência obrigatória (1.000 carateres).

Ralph Kimball, Margy Ross, *The Data Warehouse Toolkit* (3^a ed.). Wiley 2013. ISBN: 978-0471200246

Ben Sharma, *Architecting Data Lakes* (2^a ed.). O'Reilly 2018; ISBN: 978-14191952597

Bill Inmon, Mary Levins, Ranjeet Srivastava, *Building the Data Lakehouse*. Technics Publications 2021; ISBN: 978-1634629669

Salvador García, Julián Luengo, Francisco Herrera, *Data Preprocessing in Data Mining*. Springer 2015; ISBN:978-3319102474

Krish Krishnan, *Data Warehousing in the Age of the Big Data*. Morgan Kaufmann Publishers Inc. 2013. ISBN: 978-0-12-405891-0.DOI: [10.1016/C2012-0-02737-8](https://doi.org/10.1016/C2012-0-02737-8).

Alex Gorelik, *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reiley 2019. ISBN: 978-1491931554

¹ Anual, semestral, trimestral, ...

² Número total de horas de trabalho.

³ Discriminadas por tipo de metodologia adotado (T - Ensino teórico; TP - Ensino teórico-prático; PL - Ensino prático e laboratorial; TC - Trabalho de campo; S - Seminário; E - Estágio; OT - Orientação tutorial; O - Outro).

⁴ Assinalar sempre que a unidade curricular seja optativa.