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

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular ***(1.000 carateres).***
Mineração de Dados em Larga Escala / Big Data Mining

1.2. Sigla da área científica em que se insere ***(100 carateres).***
INF

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: 43,5H TP: 9H ; PL: 15H)

1.6. ECTS ***(100 carateres).***
6

1.7. Observações[4] ***(1.000 carateres).***
Optativa

1.7. Remarks ***(1.000 carateres).***
Optional

2. Docente responsável e respetiva carga letiva na Unidade Curricular ***(preencher o nome completo) (1.000 carateres).***
Nuno Miguel Soares Datia (45h)

3. Outros docentes e respetivas cargas letivas na unidade curricular ***(1.000 carateres).***
Artur Jorge Ferreira (13,5h)
Matilde Pós-de-Mina Pato (9h)

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

Os estudantes que terminam com sucesso esta unidade curricular serão capazes de:
1. Caracterizar os desafios de processar e analisar grandes volumes de dados
2. Aplicar modelos de programação e *frameworks* para processamento de dados
3. Conhecer e aplicar técnicas de redução de dimensionalidade em conjuntos de dados
4. Conhecer e aplicar técnicas de amostragem
5. Conhecer e aplicar técnicas de manipulação de dados em *streaming*
6. Conhecer e aplicar algoritmos de mineração de dados em larga escala
7. Avaliar a qualidade dos modelos produzidos e dos resultados obtidos nas tarefas de mineração
8. Interpretar soluções existentes para a mineração de dados em diferentes domínios
9. Escrever relatórios técnicos e elaborar apresentações técnicas com análise comparativa e crítica de diferentes abordagens para um dado problema

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

Students who successfully complete this course unit will be able to:
1. Characterize the challenges of processing and analysing large volumes of data
2. Apply programming models and frameworks for data processing
3. Know and apply dimensionality reduction techniques to data sets
4. Know and apply sampling techniques
5. Know and apply techniques to manipulate streaming data
6. Know and apply large-scale data mining algorithms

7. Evaluate the quality of the models produced and the results obtained in the mining tasks
8. Understand existing solutions for data mining in different domains
9. Write technical reports and prepare technical presentations with comparative and detailed analysis of different approaches for a given problem

5. Conteúdos programáticos ***(1.000 carateres).***

I. Conceito de *big data* e o fenómeno de *data deluge*. 3 V's e desafios na gestão de dados com estas caraterísticas.
II. Metodologias de programação e *frameworks* para processamento de grandes volumes de dados de forma paralela e distribuída.
III. Representação de dados. Redução de dimensionalidade: seleção e discretização de caraterísticas para aprendizagem supervisionada e não supervisionada.
IV. Manipulação de instâncias usando técnicas de amostragem probabilísticas e não probabilísticas. Subamostragem, sobre-amostragem e instâncias sintéticas.
V. Algoritmos de mineração de dados para grandes volumes de dados, para tarefas de classificação, agrupamento, associação, regressão e de recomendação.
VI. Análise de dados em *streaming*. Uso de janelas de processamento. Amostragem, sumarização, filtragem, estimativa de frequências e contagem.
VII. Análise e avaliação de dados e de modelos. Critérios de avaliação de modelos e da qualidade da aprendizagem. Análise de significância estatística.

5. Syllabus ***(1.000 characters).***

I. Concept of big data and the phenomenon of data deluge. 3 Vs and challenges manipulating datasets having such characteristics.
II. Programming methodologies and frameworks for processing large volumes of data using parallel and distributed techniques.
III. Data representation and dimensionality reduction: feature selection and discretization techniques for supervised and unsupervised learning.
IV. Instance manipulation with probabilistic and non-probabilistic sampling approaches. Subsampling, over-sampling and generation of synthetic instances.
V. Data mining algorithms for large datasets for classification, association, regression, estimation, prediction of numerical values and recommendation tasks.
VI. Streaming data analysis. Sampling, summarization, filtering, frequency estimation, and counting.
VII. Analysis and evaluation of data and models. Model evaluation metrics and the quality of learning. Analysis of statistical significance.

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

A realização de 1 trabalho prático e da componente teórica individual permitem aferir o cumprimento dos objetivos de aprendizagem (1) a (7). Com o acompanhamento, por parte do docente, da realização de cada trabalho prático, da elaboração do respetivo relatório técnico, e da apresentação do trabalho para a turma, são aferidos os objetivos de aprendizagem (8) e (9).

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

The practical assignment and individual theoretical component allow to verify the fulfilment of the learning outcomes (1) to (7). With the follow-up by the teacher during the accomplishment of each laboratory project, the writing of the corresponding technical report and the discussion of the experimental results in the classroom, it is possible to assess the fulfilment of the learning objectives (8) and (9).

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

Metodologia de ensino é teórico-prática, 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.
Os objetivos de aprendizagem de (1) a (6) são avaliados através da componente teórica, constituída por avaliação presencial (e.g. teste escrito, apresentação, e/ou teste oral) e por um resumo estendido.
Os objetivos de aprendizagem (1) a (9) são avaliados através da componente prática, que consiste na realização de um trabalho prático ao longo do semestre, escrita do respetivo relatório, e discussão oral sobre ambos.

A classificação final é obtida através de 50% da classificação da componente teórica + 50% da classificação da componente prática.
Para ambas as componentes teórica e prática, o estudante deverá obter classificação mínima de 10 valores, para obter aprovação à UC.

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

The theoretical-practical teaching is based on the Problem-Based Learning (PBL) approach. It is intended to encourage the student's autonomy in the development of solutions to complex problems, suitable to their cognitive level. Workgroup and discussion / reflection are encouraged in group sessions.
The learning outcomes (1) to (6) are evaluated through the theoretical component, consisting of face-to-face assessment (e.g. written exam, presentation, and / or oral exam) and an extended abstract.
The learning outcomes (1) to (9) are assessed through the laboratory component, which consists on one practical assignment developed along the semester, writing of the corresponding report and their oral discussion.
The final classification is the arithmetic mean of the of the theoretical component and the of the laboratory component, both with the same weight.
For both theoretical and laboratory components, the student must achieve a minimum grade of 10 points to achieve approval at the UC.

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 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).***

Theoretical lectures are provided to present the theoretical bases of the syllabus contents, using an interactive presentation of topics to help students to understand the learning outcomes. In some classes, small projects are designed and developed (theoretical-practical classes).
Autonomous work (extra-class) is guided by the laboratory work guide, designed to consolidate the skills of design and development of the learning outcomes. The project is delivered to the students in the beginning of the semester, guiding the examples and the themes lectured. All guides have a clear identification of the learning outcomes.

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

Handbook of Big Data Technologies, Albert Y. Zomaya, Sherif Sakr, Springer 2017, ISBN: 978-3319493398
Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman, Cambridge Univ. Press 2014, 2nd edition, ISBN: 978-1107015357
Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank, Mark A. Hall, Morgan Kaufmann Publishers 2016, 4th edition, ISBN: 978-0128042915

---

[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.