

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

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular (1.000 carateres).

Técnicas estatísticas para análise de mineração de dados / Statistical techniques for data mining analysis

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

MAT

1.3. Duração¹ (100 carateres).

Semestral

1.4. Horas de trabalho² (100 carateres).

160

1.5. Horas de contacto³ (100 carateres).

TP: 45 PL:22.5 OT: 5

1.6. ECTS (100 carateres).

6

1.7. Observações⁴ (1.000 carateres).

1.7. Remarks (1.000 carateres).

2. Docente responsável e respetiva carga letiva na Unidade Curricular (preencher o nome completo) (1.000 carateres).

Sandra Maria da Silva Figueiredo Aleixo (27h)

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

Iola Maria Silvério Pinto (27h)

Gonçalo Caetano Marques(13.5h)

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

Esta UC introduz os algoritmos e métodos estatísticos para mineração de dados. A sua natureza interdisciplinar combina tópicos de estatística, bases de dados e ciência da computação. Os objetivos são:

1. Identificar as fases de um projeto de Data Science. Conhecer os conceitos e técnicas estatísticas de aprendizagem automática (AA), exemplificando as suas aplicações e funcionalidades
2. Identificar os vários tipos de dados de input e conhecer os métodos para a sua preparação e pré-processamento, bem como as principais técnicas para transformação e redução da dimensionalidade dos dados
3. Conhecer os fundamentos teóricos matemáticos dos métodos de AA apresentados, saber manuseá-los, bem como identificar e interpretar várias formas e tipos de resultados
4. Avaliar os resultados obtidos com as técnicas de AA, usando e interpretando as medidas de desempenho
5. Usar softwares apropriados para resolver vários desafios
6. Realizar na íntegra um projeto de Data Science usando metodologias adequadas

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

This CU introduces statistical data mining algorithms and methodologies. It's interdisciplinary nature combines topics of statistics, databases and computer science. The intended outcomes are:

1. To describe the several stages of a Data Science project. To know the concepts and statistical techniques of machine learning (ML), exemplifying its applications and functionalities
2. To identify the several types of input data and to know the methodologies for its preparation and preprocessing, as well as the main techniques for transformation and reduction of data dimensionality

- 3 To know the mathematical theoretical foundations associated to the ML methods, to know how to handle them, as well as to identify and interpret the several forms and types of results
4. To evaluate the results obtained with the ML techniques, using and interpreting the performance measures
5. Use appropriate softwares for solving several challenges
6. To complete a Data Science project using appropriate methodologies

5. Conteúdos programáticos (1.000 carateres).

1. Introdução. Fases de um projeto de Data Science. Conceitos e ferramentas de mineração de dados e aprendizagem automática. Input: tipos de dados e preparação de dados. Output: tipos de representação
2. Transformações de dados. Métodos de redução de dimensionalidade.
3. Modelos e algoritmos de classificação, regressão e agrupamento
4. Avaliação e comparação de modelos e algoritmos

5. Syllabus (1.000 characters).

1. Introduction. Stages of a Data Science project. Data mining and machine learning concepts and tools. Input: data types and data preparation. Output: types of representation
2. Data transformations. Dimensionality reduction methods
3. Models and algorithms of classification, regression and clustering
4. Evaluation and comparison of models and algorithms

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

O ponto 1 dos conteúdos programáticos pretende alcançar os pontos 1 e 2 dos objetivos

O ponto 2 dos conteúdos programáticos introduz os conceitos necessários para atingir o ponto 2 dos objetivos

O ponto 3 dos conteúdos programáticos pretende alcançar o ponto 3 dos objetivos

O ponto 4 dos conteúdos programáticos concentra-se na consolidação dos objetivos 1 e 4

Os objetivos referidos no ponto 5 e 6 são alcançados em todos os pontos dos conteúdos programáticos

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

Point 1 of the syllabus aims to achieve the points 1 and 2 of the objectives

Point 2 of the syllabus introduces the concepts necessary to achieve the point 2 of the objectives

Point 3 of the syllabus aims to achieve the point 3 of the objectives

Point 4 of the syllabus focus on the consolidation of goals 1 and 4

The objectives referred in the goal 5 and 6 are achieved throughout all points of the syllabus

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

As aulas são teórico-práticas. É utilizada uma metodologia expositiva para a apresentação da matéria teórica, exemplificada com a resolução de exercícios e de problemas concretos, implementada computacionalmente usando softwares adequados, nomeadamente R, WEKA e Python.

A avaliação de conhecimentos compreende duas componentes, uma teórica (NT) e outra prática (NP). A componente teórica é constituída por um exame (nota mínima de 9,5 valores). A componente prática é constituída por um trabalho de grupo (nota mínima de 9,5 valores). Este trabalho será um dos casos de estudo apresentado de entre vários, deve ser desenvolvido ao longo do semestre e deve integrar as várias fases de um projeto de Data Science.

A nota final do aluno (NF) será obtida através da fórmula $NF=0,4NT+0,6NP$.

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

The classes are theoretical-practical. An expository methodology is used for the presentation of the theoretical matter, exemplified with the resolution of exercises and concrete problems, implemented in a computational way using suitable softwares, namely, R, WEKA and Python.

The knowledge assessment comprises two components, one theoretical (TG) and another practical (PG). The theoretical component consists of an exam (minimum grade of 9.5 points). The practical component consists of a group work (minimum grade of 9.5 values). This work will be one of the case studies presented among several, should be developed throughout the semester and should integrate the various phases of a Data Science project.

The student's final grade (FG) will be obtained through the formula $FG = 0.4TG + 0.6PG$.

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

As metodologias de ensino são coerentes com os objetivos da unidade curricular, dado que a metodologia expositiva

utilizada para explicar a matéria teórica possibilita atingir os objetivos da unidade curricular. A utilização de exemplos resolvidos computacionalmente associados aos diversos tópicos do programa, permite dotar os alunos de competências adequadas para a resolução dos desafios colocados pelo tecido empresarial no mercado de trabalho. A elaboração de um projeto completo de Data Science, usando metodologias apropriadas, será uma mais valia para quando os alunos iniciarem a sua atividade profissional.

O método de avaliação permite averiguar se o aluno adquiriu os conhecimentos necessários para atingir os objetivos propostos na unidade curricular.

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

The teaching methodologies are consistent with the objectives of the curricular unit, given that the expository methodology used to explain the theoretical subject makes it possible to reach the objectives of the curricular unit. The use of computationally solved examples, associated with the various topics of the program allows students to have adequate skills to solve the challenges posed by the job market. The implementation of a complete Data Science project, using appropriate methodologies, will be an advantage when students begin their professional activity.

The evaluation method allows to verify if the student has acquired the necessary knowledge to reach the objectives proposed in the curricular unit.

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

Principal

1. Bishop, C. M., Pattern Recognition and Machine Learning, Springer (2006).
2. Duda, R. O., Hart, P. E. and Stork, D. G., Pattern Classification, Wiley (2001).
3. Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. Springer (2017).
4. Lantz, B., Machine Learning with R, Packt (2013).
5. Muller, A. and Guido, S., Introduction to Machine Learning with Python, O'Reilly (2017)
6. Murphy, K., Machine Learning: A Probabilistic Perspective, MIT Press (2012).
7. Witten, I. H., Frank, E. and Hall M. A., Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2011).

Secundária

8. Hand, D. J., Mannila, H. and Smyth, P., Principles of data mining. The MIT Press (2001).
9. Raschka, S., Python machine learning. Packt Publishing (2015).
10. Rocha, M., Cortez, P. and Neves, J., Análise inteligente de dados – algoritmos e implementação em Java. FCA (2008).
11. Shmueli, G., Bruce, P. C. and Patel, N.R., Data mining for business analytics. Wiley (2016)
12. Torgo, L., Data mining with R - learning with case studies. CRC Press (2010).

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.