
1. Designação da unidade curricular

[4236] Técnicas Estatísticas para Análise de Mineração de Dados / Statistical Techniques for Data Mining Analysis

2. Sigla da área científica em que se insere

MAT

3. Duração

Unidade Curricular Semestral

4. Horas de trabalho

162h 00m

5. Horas de contacto

Total: 67h 30m das quais TP: 45h 00m | P: 22h 30m

6. % Horas de contacto a distância

Sem horas de contacto à distância

7. ECTS

6

8. Docente responsável e respetiva carga letiva na Unidade Curricular

[1409] Iola Maria Silvério Pinto | Horas Previstas: 67.5 horas

9. Outros docentes e respetivas cargas letivas na unidade curricular

Não existem docentes definidos para esta unidade curricular

10. Objetivos de aprendizagem e a sua compatibilidade com o método de ensino (conhecimentos, aptidões e competências a desenvolver pelos estudantes).

A UC introduz algoritmos e métodos estatísticos para mineração de dados. Alia estatística, bases de dados e computação. Os objetivos são:

1. Rever conceitos de Probabilidade e Estatística. Identificar as fases de um projeto de Ciência de Dados (PCD). Aprender conceitos e técnicas estatísticas de aprendizagem automática (AA)
2. Identificar o tipo de dados e aprender métodos para a sua preparação e pré-processamento
3. Planear e implementar uma base de dados segundo o modelo relacional
4. Aplicar técnicas para transformação no âmbito dos modelos de regressão nomeadamente modelos lineares generalizados, modelos aditivos generalizados e redes neurais
5. Conhecer a base matemática dos métodos de AA apresentados, saber manuseá-los, identificar e interpretar várias formas e tipos de resultados
6. Avaliar os resultados obtidos com as técnicas de AA, usando e interpretando as medidas de desempenho
7. Usar softwares apropriados
8. Realizar um PCD no âmbito da regressão, com metodologias adequadas

10. Intended Learning objectives and their compatibility with the teaching method (knowledge, skills and competences by the students).

The CU introduces statistical data mining algorithms and methodologies. It combines statistics, databases and computation. The intended outcomes are:

1. Review concepts of Probability and Statistics. To describe the several stages of a Data Science project (DSP). To know the concepts and statistical techniques of machine learning (ML)
2. Identify the type of data and to know the methodologies for its preparation and preprocessing
3. Planning and implement a data base accordingly to the relational model
4. Apply techniques for transformation in the regression models context, such as generalized linear model, generalized additive model and neural networks
5. Know the mathematical basis of the ML methods presented, know how to handle them, identify and interpret various forms and types of results
6. Evaluate the results obtained with the ML techniques, using and interpreting the performance measures
7. Use appropriate software
8. Complete a DSP under the regression, with appropriate methodologies

11. Conteúdos programáticos

1. Revisão de Probabilidade e Estatística.
2. Introdução à Mineração de Dados. Projeto de Mineração de Dados.
3. Dados: Tipos; Qualidade; Pré-processamento.
4. Bases de Dados: Modelo Entidade Relacionamento; Implementação do Modelo Físico; Consultas.
5. Métodos Lineares para Regressão - Ridge e Lasso.
6. Modelos Lineares Generalizados. Modelos de Regressão Logística.
7. Análise de Sobrevivência.
8. Métodos Não Lineares para Regressão.
9. Modelos Aditivos Generalizados.

11. Syllabus

1. Review of Probability and Statistics.
2. Introduction to Data Mining. Data Mining Project.
3. Data: Type; Quality; Pre-processing.
4. Data Bases: The Entity-Relationship Model; Physical Model implementation; Queries.
5. Regression Linear Methods - Ridge and Lasso.
6. Generalized Linear Models. Logistic Regression Models.
7. Survival Analysis.
8. Non Linear Regression Methods.
9. Generalized Additive Models.

12. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular

Os pontos 1 e 2 dos conteúdos programáticos pretendem alcançar o ponto 1 dos objetivos

O ponto 3 dos conteúdos programáticos introduz os conceitos necessários para atingir o ponto 2 dos objetivos

O ponto 4 dos conteúdos programáticos pretende alcançar o ponto 3 dos objetivos

Os restantes pontos (5 a 9) dos conteúdos programáticos pretendem alcançar os restantes objetivos (4 a 8)

12. Evidence of the syllabus coherence with the curricular unit's intended learning outcomes

Topics 1 and 2 of the syllabus aims to achieve the topic 1 of the objectives

Topic 3 of the syllabus introduces the concepts necessary to achieve the objectives topic 2

Topic 4 of the syllabus aims to achieve the topic 3 of the objectives

The rest of syllabus topics (5 until 9) focuses on the consolidation of goals 4 until 8

13. Metodologias de ensino e de aprendizagem específicas da unidade curricular articuladas com o modelo pedagógico

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 um software estatístico. Desde que reunidas as condições necessárias a UC pode ser parcialmente lecionada à distância de forma síncrona (1/3 das horas de contacto semanais).

13. Teaching and learning methodologies specific to the curricular unit articulated with the pedagogical model

Classes are theoretical and practical. Expository methodology is used for the presentation of the theoretical subjects, exemplified with the resolution of exercises and case-studies, implemented in a computational way, supported by a statistical software. As long as the necessary conditions are met, this CU can be partially taught remotely in a synchronous manner (1/3 of the weekly contact hours).

14. Avaliação

A avaliação de conhecimentos é distribuída, com exame final, compreende duas componentes, uma teórica (NT) e outra prática (NP). A componente teórica é constituída por um exame escrito com nota mínima de 9.5 valores. A componente prática é constituída por um trabalho pedagogicamente fundamental, a realizar em grupo, com a nota mínima de 9.5 valores, com entregas parciais distribuídas ao longo do semestre, e com apresentação e discussão obrigatória para todos os membros do grupo.

A nota final do aluno, NF, será obtida através da fórmula $NF=0.5NT+0.5NP$.

14. Assessment

The knowledge assessment is distributed, with final exam, comprises two components, one theoretical (TG) and another practical (PG). The theoretical component consists of a written exam (minimum grade of 9.5 points). The practical part consists of a pedagogically fundamental group work, with a minimum grade of 9.50, including partial deliveries distributed throughout the semester and with mandatory presentation and discussion for all group members.

The student's final grade (FG) will be obtained through the formula $FG = 0.5TG + 0.5PG$.

15. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular

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 Ciência de Dados, 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.

15. Evidence of the teaching methodologies coherence with the curricular unit's intended learning outcomes

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.

16. Bibliografia de consulta/existência obrigatória

1. James, G., Witten, D., Hastie, T. and Tibshirani, R., An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics (2017).
2. Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. Springer (2017).
3. Bishop, C. M., Pattern Recognition and Machine Learning, Springer (2006)
4. Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V., Introduction to Data Mining, Pearson (2019)
5. Witten, I. H., Frank, E. and Hall M. A., Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2011).
6. Hand, D. J., Mannila, H. and Smyth, P., Principles of data mining. The MIT Press (2001).
7. Torgo, L., Data mining with R ? learning with case studies. CRC Press (2010).
8. Zhao, Y., R and Data Mining: Examples and Case Studies, Elsevier (2012).
9. Lantz, B., Machine Learning with R, Packt (2013).
10. Muller, A. and Guido, S., Introduction to Machine Learning with Python, O'Reilly (2017).

17. Observações

Unidade Curricular Obrigatória

Unidade Curricular comum ao(s) curso(s) de PGMCD

Data de aprovação em CTC: Data de aprovação em CTC: 2024-07-17

Data de aprovação em CP: Data de aprovação em CP: 2024-06-26