Universidade do Minho
Departamento de Informática
Mestrado Integrado em Engenharia Informática
Mestrado em Engenharia Informática

Perfil Sistemas Inteligentes
Aprendizagem e Extracção de Conhecimento
1º/4º Ano, 1º Semestre
Ano letivo 2020/2021

Trabalho prático – 2ª Fase
Novembro de 2020

Tema	APRENDIZAGEM E EXTRAÇÃO DE CONHECIMENTO
Objetivos de aprendizagem	Com a realização deste trabalho prático pretende-se que os alunos apliquem os conhecimentos adquiridos na unidade curricular de Aprendizagem e Extração de Conhecimento: • Análise e preparação de conjunto de dados; • Desenvolvimento de modelos preditivos; • Validação e otimização da *performance* dos modelos preditivos desenvolvidos.
Enunciado	Este enunciado pretende ser o ponto de partida para o desenvolvimento de um modelo de classificação utilizando o ambiente de desenvolvimento Python, aplicando as funcionalidades disponíveis na biblioteca Sklearn. Como tal, será necessário o desenvolvimento de uma solução para o seguinte problema: *Preparação e análise de um dataset relativo às características de funcionários de múltiplas empresas, como forma de prever o nível salarial anual do indivíduo.* Este projeto baseia-se num conjunto de casos de estudo apresentando informações referentes a funcionários de empresas espalhadas pelo mundo, onde são analisadas múltiplas características sobre o indivíduo como forma de prever a sua classe salarial anual. Anexo a este trabalho prático encontram-se dois ficheiros: • *training.csv*, onde se apresentam os casos de estudo a serem aplicados exclusivamente para o **treino** do modelo preditivo; • *test.csv*, onde se apresentam os casos de estudo a serem aplicados exclusivamente para a **análise** e **validação** do modelo preditivo; • *attribute_info.docx,* onde são apresentados alguns detalhes relativamente aos atributos do conjunto de dados fornecidos. Para a resolução do problema deve começar por analisar, preparar e visualizar a distribuição do *dataset*, de modo a interpretar a sua relação com a variável a prever (i.e., salary-classification). Tendo em conta esta análise, o processo seguinte passa por desenvolver diferentes modelos de classificação e validar a sua performance. No final, o grupo deverá selecionar o modelo que apresenta melhor performance de classificação e justificar a sua decisão. Como métrica de avaliação da performance do modelo classificador, deverá ser aplicada a métrica de validação **acurácia**. Como forma de justificar este resultado, deverá apresentar a sua respetiva matriz de classificação.

	Este trabalho prático compreende a entrega do código desenvolvido e do relatório, em formato digital, apresentando de forma sucinta todos os procedimentos aplicados e respetivas justificações da sua utilização, apoiando-se na demonstração dos resultados adquiridos.
Entrega	A data para a entrega final do relatório e apresentação dos respetivos algoritmos desenvolvidos é fixada no dia 10 de janeiro de 2021. A sessão de apresentação decorrerá no período de aulas correspondente desta unidade curricular nos dias 11 e 12 de janeiro de 2021. O código resultante da realização do trabalho prático e o respetivo relatório em formato digital .PDF deverão ser submetidos através da página de submissão disponível na pasta da UC "Conteúdo/Instrumentos de Avaliação/Trabalho prático (2ª parte): Sistemas de Aprendizagem", em ficheiros compactados (formato ZIP). O ficheiro deverá ser identificado na forma "AEC_F2GXX", em que [XX] designa o número do grupo de trabalho. O documento deverá seguir as instruções apresentadas para a coleção LNCS @ Springer, em formato artigo científico Cada grupo disporá de 10 minutos para a apresentação dos principais resultados alcançados.
Referências bibliográficas	Bowles, M. (2015). *Machine learning in Python: essential techniques for predictive analysis*. John Wiley & Sons. Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.