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 2019/2020

Trabalho prático – 2ª Fase
Novembro, 2019

Tema	APRENDIZAGEM E EXTRAÇÃO DE CONHECIMENTO
Objetivos de aprendizagem	Com a realização deste trabalho prático pretende-se que os alunos aprendam os seguintes procedimentos utilizados em Projetos de Extração de Conhecimento: • Preparação e estruturação de *features* e *target variables* para treino de modelos de Extração de Conhecimento; • Treino/Validação de modelos preditivos de Machine Learning; • Análise 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/Sklearn. Para isso, será necessário o desenvolvimento de uma solução para o seguinte problema: *Preparação e análise de um dataset relativo às condições de funcionários de uma empresa, como forma de prever a sua ausência.* Este projecto baseia-se num conjunto de casos de estudo apresentando informações referentes a funcionários de uma empresa de transporte no Brasil, onde são analisados os motivos associados à ausência ao trabalho entre Julho de 2007 e Julho de 2010. Este conjunto de dados foram extraídos do Repositório de Aprendizado de Máquina da UCI (http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work). O conjunto de dados possui 740 casos de estudo com 20 features distintas (que podem ou não estar relacionados à ausência do trabalho) adquiridas de 36 funcionários diferentes. Anexo a este trabalho prático encontram-se 4 ficheiros: (1) *train_data.csv*, onde apresenta os casos de estudo a serem aplicados exclusivamente para o **treino** do modelo preditivo; (2) *test_data.csv*, onde apresenta os casos de estudo a serem aplicados exclusivamente para a **análise** e **validação** do modelo preditivo; (3) *sample_submission.csv*, onde apresenta os resultados do target variables associado aos casos de estudo apresentados no fich. *test_data.csv*. Como métrica de avaliação da performance do modelo classificador, deverá ser aplicado a métrica de score **Accuracy.** Definido o problema, faz-se uma breve explicação de cada uma das *features*: • ID – identificador único do funcionário; • Reason for absence – número identificador associado à razão da ausência estruturadas pelo Código Internacional de Doenças (CID) – info. adicional disponível no fich. em anexo "Attribute Information.docx"; • Month of absence – mês do caso de estudo; • Day of the week – dia da semana do caso de estudo (segunda-feira (2), terça-feira (3), quarta-feira (4), quinta-feira (5), sexta-feira (6)); • Seasons – estação do ano do caso de estudo; • Transportation expense – gastos financeiros associados ao transporte do funcionário; • Distance from Residence to Work – distancia entre o local de trabalho e a residência do funcionário (quilómetros);

- Service time – tempo de serviço anual que o funcionário apresenta na empresa;
- Age – idade do funcionário;
- Work load Average/day – carga de trabalho médio que o funcionário apresenta por dia;
- Hit target – carga de trabalho percentual concluído pelo funcionário;
- Disciplinary failure – define se o funcionário já apresenta pelo menos uma falha disciplinar dentro da empresa (sim=1; não=0);
- Education – grau de escolaridade do funcionário (ensino médio (1), licenciatura (2), pós-graduação (3), mestrado e doutorado (4));
- Son – número de filhos que o funcionário apresenta;
- Pet – número de animais de estimação que o funcionário apresenta;
- Social drinker – define se o funcionário é bebedor social (sim=1; não=0);
- Social smoker – define se o funcionário é fumador (sim=1; não=0);
- Weight – peso do funcionário;
- Height – altura do funcionário;
- Body mass índex – índice de massa corporal do funcionário.

Através deste conjunto de dados, o *target variable* a prever é apresentado pelo seguinte dado:

- Absent – apresenta se o funcionário irá ausentar ao trabalho (sim=1; não=0).

Para a resolução do problema deve começar por analisar e visualizar a distribuição das *features* do *dataset*, de modo a interpretar a relação com o *target variable* (i.e., Absent). Com este conhecimento, técnicas de tratamento de dados (verificação de outliers, normalização de dados, categorização de features, selecção de features mais relevantes, etc.) deverão ser aplicadas, como forma de melhorar a *performance* de classificação do modelo preditivo. Este processo terá em conta o treino e validação dos diferentes modelos preditivos leccionados na unidade curricular, como forma de comparar a performance de classificação dos respectivos modelos (designado benchmark). Como extra, técnicas de *hyperparameterization optimization* poderão ser aplicadas como forma de optimizar o processo de classificação dos modelos preditivos.

Esta 2ª fase do trabalho prático compreende a entrega do código desenvolvido e do relatório, definindo todos os procedimentos aplicados e respectivas justificações da sua utilização, apoiando-se na demonstração dos resultados adquiridos.

Entrega

A data para a entrega final do relatório e apresentação dos respectivos algoritmos desenvolvidos é fixada no dia 15 de dezembro de 2019. A sessão de apresentação decorrerá no período de aulas correspondente desta unidade curricular no dia 16 de dezembro de 2019.

O código resultante da realização do trabalho prático e o respetivo relatório em formato digital .PDF deverão ser enviados por correio eletrónico para pjon@di.uminho.pt e fgoncalves@algoritmi.uminho.pt, em ficheiros compactados (formato ZIP). Tanto o assunto da mensagem como o ficheiro deverão ser identificados na forma "[AEC: FYGXX]", em que [Y] designa o número da fase e [XX] designa o número do grupo de trabalho.

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.

Trivedi, H., *Explaining Absenteeism at Workplace Predicted by a Neural Network*.