



Universidade do Minho

Departamento de Informática

Mestrado [integrado] em Engenharia Informática

Perfil de Machine Learning: Fundamentos e Aplicações

Sistemas Baseados em Similaridade

4º/1º Ano, 1º Semestre

Ano letivo 2020/2021

Trabalho Prático de Grupo – 1ª Parte

Outubro, 2020

<b>Tema</b>	Conceção e implementação de modelos de <i>Machine Learning</i> baseados em árvores.
<b>Objetivos de Aprendizagem</b>	Com a realização deste trabalho prático pretende-se sensibilizar e motivar os alunos para a conceção e desenvolvimento de um projeto de <i>Machine Learning</i> utilizando a plataforma <i>Knime</i> e modelos baseados em árvores, como Árvores de Decisão ou <i>Random Forests</i> .
<b>Enunciado</b>	<p>A 1ª parte da componente prática de avaliação de grupo incidirá sobre as seguintes tarefas:</p> <ul style="list-style-type: none"><li>• Os grupos de trabalho deverão consultar, analisar e selecionar um <i>dataset</i> de entre os que estão acessíveis a partir de fontes como, por exemplo, o <i>Google Dataset Search</i> ou <i>Kaggle</i>;</li><li>• Para além do <i>dataset</i> selecionado no ponto anterior, os grupos deverão também trabalhar o <i>dataset</i> disponível em <a href="https://www.kaggle.com/c/20tp2sbs21">https://www.kaggle.com/c/20tp2sbs21</a>:<ul style="list-style-type: none"><li>○ O link anterior redireciona para a plataforma <i>Kaggle</i> onde foi criada uma competição. O <i>dataset</i> a utilizar na competição, assim como todos os detalhes do funcionamento da mesma, estão disponíveis no link acima referido. Em suma, deverão conceber e treinar um modelo, aplicá-lo sobre um conjunto de dados de teste e submeter os resultados na plataforma, a qual avaliará a <i>accuracy</i> do modelo desenvolvido;</li><li>○ Na plataforma <i>Kaggle</i> deverão formar equipas com os restantes elementos do grupo de trabalho. O nome da equipa deverá seguir o formato <b>GRUPO_X_MLFA</b> onde <b>X</b> corresponde ao número do grupo;</li><li>○ Deverão utilizar o seguinte link para se inscreverem na competição: <a href="https://www.kaggle.com/t/8807337a10e040b69d0bb63356327cbe">https://www.kaggle.com/t/8807337a10e040b69d0bb63356327cbe</a></li></ul></li><li>• Utilizar a plataforma <i>Knime</i> para desenvolver um, ou mais, <i>workflows</i> para:<ul style="list-style-type: none"><li>○ Exploração, análise e preparação dos dados de ambos os <i>datasets</i>;</li><li>○ Extração de conhecimento dos dados;</li><li>○ Conceber e otimizar modelos baseados em árvores;</li><li>○ Obtenção e análise crítica de resultados.</li></ul></li><li>• Interpretar os resultados e a sua utilidade no contexto dos problemas subjacentes aos <i>datasets</i> trabalhados. Determinar quais os resultados mais relevantes;</li><li>• Criação de objetos visuais que permitam ter uma noção gráfica dos dados, dos modelos e dos resultados obtidos;</li></ul>

- Submeter os resultados obtidos no *Kaggle* de forma a obter a *accuracy* do modelo, sendo que existe um limite diário de 3 submissões válidas. Este passo refere-se, obviamente, apenas ao *dataset* da competição.

Os resultados obtidos deverão ser objeto de um relatório que contenha, entre outros:

- Quais os domínios a tratar, quais os objetivos e como se propõe a atingi-los;
- Qual a metodologia seguida e como foi aplicada;
- Descrição e exploração detalhada do *dataset* e de todo e qualquer tratamento efetuado ao mesmo;
- Descrição dos *workflows* criados e com que objetivo (não se pretende uma descrição nodo-a-nodo). Quais os principais nodos e como foram configurados, entre outros detalhes que seja oportuno fornecer;
- Descrição detalhada do modelo desenvolvido, quais as suas características, como e sobre que parâmetros foi realizado o *tuning* do modelo, características do treino, entre outros detalhes que seja oportuno fornecer;
- Sumário dos resultados obtidos e respetiva análise crítica;
- Apresentação de sugestões e recomendações após análise dos resultados obtidos e dos modelos desenvolvidos.

Todo o processo deverá ser acompanhado de exemplos e indicações que permitam reproduzir todos os passos realizados assim como os resultados obtidos.

## Entrega e Avaliação

As submissões deverão ser feitas na plataforma de e-learning da Universidade do Minho, em “*Conteúdo/Instrumentos de Avaliação em Grupo/Submissão TP1*”, enviando, num único ficheiro compactado, todos os elementos produzidos. O ficheiro submetido deverá estar identificados na forma “SBS\_TP1\_GX”, em que X designa o número do grupo de trabalho. As submissões deverão ser realizadas até ao final do dia 25 de novembro de 2020.

A sessão de apresentação do trabalho decorrerá no dia 26 de novembro de 2020, no auditório 0.08 do CP1, tendo início às 13h00min. Cada grupo disporá de 10 minutos para realizar a apresentação, utilizando os meios que considerar mais adequados.