Universidade do Minho
Departamento de Informática
Mestrado em Matemática e Computação
Mestrado em Engenharia de Sistemas

Perfil de Machine Learning: Fundamentos e Aplicações
Sistemas Baseados em Semelhança
1º/2º Ano, 1º Semestre
Ano letivo 2023/2024

Enunciado Prático nº 3
12 de outubro de 2023

Tema

A Data Science Perspective

Enunciado

Uma exploração aprofundada dos dados permite que se tirem ilações, muitas vezes escondidas, que poderão ser importantes para se compreender o domínio e o problema em mãos. Com este enunciado prático é esperado que sejam aplicadas um conjunto de técnicas que permitam explorar e tratar *datasets*.

Tarefas

Numa primeira fase devem descarregar o *dataset* disponível em https://goo.gl/p2y19t que contém dados de um conjunto de utilizadores de uma determinada plataforma web assim como o seu sentimento em relação à mesma. Devem, de seguida:

T1. Carregar, no *Knime*, o *dataset* descarregado. Aplicar nodos para exploração de dados, i.e., analisar os dados em relação à sua:

a. Tendência central;
b. Dispersão estatística;
c. Correlação entre *features*.

T2. Criar *plots* para visualização dos dados;

T3. Aplicar nodos para tratamento de dados de forma a:

a. Excluir todas as colunas do tipo *Double*;
b. Tratar valores em falta;
c. Remover registos duplicados;
d. Criar 3 *bins* de igual frequência para a *feature age*;
e. Para cada registo, extrair o ano, mês e dia da semana da *feature birthday*;
f. Excluir utilizadores da plataforma que tenham uma atividade na plataforma (*WebActivity*) inferior a 1 hora e que tenham mais de 70 anos;
g. Excluir todos os registos que contenham a *sub-string* "co" no produto.

T4. Aplicar nodos para agregação de dados de forma a:

a. Por género, obter o número e a percentagem de registos, assim como a média da idade e da atividade na plataforma. Obter também o mínimo e máximo da idade;
b. Por género e atividade na plataforma, obter a moda da análise do sentimento em relação à plataforma e a média da avaliação do sentimento;

c. Por análise de sentimento, obter o número de registos, a média do salário anual estimado, o somatório do salário anual e a média do número de contratos.

T5. Análise crítica à informação extraída das agregações efetuadas na tarefa anterior. Que conclusões poderia a empresa tirar?

Numa segunda fase devem descarregar o *dataset* disponível em https://bit.ly/3525yDr. Este *dataset* contém dados referentes à performance de vários jogadores de futebol na edição 2017/2018 da *Premier League*. Devem, de seguida:

T6. Carregar, no *Knime*, o *dataset* descarregado. Explorar os dados, procurar informação relevante e mostrar essa mesma informação. P.e., qual a equipa mais indisciplinada? Qual o top-10 dos assistentes para golo? Qual o top-5 de nacionalidades na liga?