Universidade do Minho
Licenciatura em Engenharia Informática

Aprendizagem e Decisão Inteligentes
3º Ano, 2º Semestre
Ano letivo 2024/2025

Guião prático nº 18
Abril, 2025

Tema

Aplicação de técnicas de aprendizagem com KNIME: Segmentação/ *Clustering*

Objetivos de aprendizagem

Com a realização desta ficha prática pretende-se que os estudantes:

- Apliquem nodos de aprendizagem não supervisionada, de segmentação;
- Usem nodos de avaliação de modelos;

Enunciado

Os dados incluídos nos ficheiros [college_data*.csv] descrevem decisões sobre a frequência de instituições de ensino privadas ou públicas (kaggle.com/datasets/faressayah/college-data).

Este problema de classificação binária (privada/pública) deve ser abordado por paradigmas de aprendizagem sem supervisão, usando técnicas de segmentação (*clustering*) com vista à aplicação dos algoritmos K-means e K-medoids.

Realizar as tarefas seguintes:

1. Carregar o *dataset* [college_data_train] e aplicar nodos de exploração de dados;
2. Proceder ao tratamento e limpeza dos dados;
3. Aplicar o nodo K-means para treinar o modelo de aprendizagem não supervisionado, de modo a classificar cada caso de estudo como "instituto privado" ou "instituto público" (*number of clusters* = 2);
4. Aplicar os nodos de visualização COLOR MANAGER, SHAPE MANAGER e SCATTER PLOT para representar os diferentes casos de estudo e respetivos *clusters* associados;
5. Carregar o *dataset* [college_data_test] que representa o conjunto de dados de teste (com a inclusão do atributo "Private", representando se a universidade é um instituto privado ou público).
    a. Proceder ao tratamento e limpeza dos dados;
6. Aplicar o nodo CLUSTER ASSIGNER para inferir sobre os dados de teste utilizando o modelo treinado no nodo K-means.
    a. Aplique o nodo RULE ENGINE para adequar o nome dos *clusters* atribuídos para cada caso de estudo à respetiva classificação do instituto (coluna "Private");
7. Avalie o desempenho dos modelos de aprendizagem K-means através do uso de matrizes de confusão e métricas de avaliação (use o nodo SCORER (JAVASCRIPT)).
    a. Quais os resultados obtidos?
    b. Em que situações o modelo acerta/falha?
    c. Como melhorar o modelo de aprendizagem proposto?

Descrição do *dataset* MOON OR NOT MOON

ATRIBUTO	DESCRIÇÃO
Institute	Nome da instituição
Private	*Target* com os valores "No" e "Yes" indicando universidade privada ou pública
Apps	Número de candidaturas
Accept	Número de inscrições aceites
Enroll	Número de novos alunos matriculados
Top10perc	Percentagem de novos alunos entre os 10% melhores da turma do ensino secundário
Top25perc	Percentagem de novos alunos entre os 25% melhores da turma do ensino secundário
F.Undergrad	Número de estudantes de licenciatura em tempo integral
P.Undergrad	Número de estudantes de licenciatura a tempo parcial
Outstate	Propina para estudantes de fora do estado
Room.Board	Custos de alojamento e alimentação
Books	Custos estimados com livros
Personal	Gastos pessoais estimados
PhD	Percentagem de docentes com doutoramento
Terminal	Percentagem de docentes com grau académico terminal
S_F_Ratio	Rácio aluno/professor
perc.alumni	Percentagem de ex-alunos que fizeram doações
Expend	Despesa de instrução por aluno
Grad.Rate	Taxa de graduação

Mais detalhes sobre estes dados podem ser encontrados neste *link*: kaggle.com/datasets/faressayah/college-data