Universidade do Minho
Licenciatura em Engenharia Informática

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

Guião prático nº 5
Fevereiro, 2025

Tema

Exploração avançada e preparação de dados com KNIME

Objetivos de aprendizagem

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

- Explorem a plataforma de análise de dados KNIME;
- Experimentem tarefas de exploração avançada de dados;
- Realizem tarefas de preparação de dados;

Enunciado

Os dados respeitam a voos e respetivos atrasos nas chegadas (*target*) de diversas companhias aéreas.

Realizar as tarefas seguintes:

1. Carregar o *dataset* «flights» com os dados dos voos;
2. Aplicar nodos para exploração de dados que respondam à resolução das seguintes situações:
    a. Em que colunas os valores são sempre iguais?
    b. Que colunas têm variância elevada?
    c. Que colunas representam o mesmo conhecimento?
    d. Existe alguma relação relevante entre as *features*?
3. Aplicar nodos de exploração e tratamento de dados para:
    a. Remover colunas que possam ser irrelevantes ou que não traduzam conhecimento novo;
    b. Verificar a existência de entradas duplicadas;
    c. Tratar valores em falta (*missing values*);
    d. Remover *outliers*;
    e. Realizar *encoding* das *features* (transformar dados em colunas);
    f. Fazer a normalização de valores;
4. Criar modelos de previsão:
    a. Treinar um modelo de regressão linear para prever o atraso na chegada;
    b. Analisar o resultado da previsão.

Descrição do *dataset* FLIGHTS

ATRIBUTO	DESCRIÇÃO
id	Identificador do voo
year	Ano em que o voo se realizou
month	Mês em que o voo se realizou (1 = janeiro; 12 = dezembro).
day	Dia em que o voo se realizou
dep_time	Hora de partida real (formato "hhmm")
sched_dep_time	Hora de partida agendada (formato "hhmm")
dep_delay	Atraso na partida do voo (em minutos) - diferença entre horário real e agendado - valores positivos são atraso e negativos são antecipação
arr_time	Hora de chegada real (formato "hhmm")
sched_arr_time	Hora de chegada agendada (formato "hhmm")
arr_delay	Atraso na chegada do voo (em minutos) - diferença entre horário real e agendado - valores positivos são atraso e negativos são antecipação
carrier	Código da companhia aérea (duas letras)
flight	Número do voo
tailnum	Identificador do avião
origin	Código do aeroporto de partida (três letras)
dest	Código do aeroporto de destino (três letras)
air_time	Duração do voo (em minutos)
distance	Distância entre os aeroportos de origem e destino (em milhas)
hour	Hora da partida agendada
minute	Minutos da partida agendada
time_hour	Hora agendada de partida do voo ("aaaa-mm-dd hh:00:00")
name	Nome completo da companhia aérea responsável pelo voo

Mais detalhes sobre estes dados podem ser encontrados neste link: kaggle.com/datasets/flights