

Universidade do Minho

Licenciatura em Engenharia Informática

Aprendizagem e Decisão Inteligentes 3° Ano, 2° Semestre Ano letivo 2023/2024

Ficha prática nº 2 Fevereiro, 2024

Tema

Exploração de dados com KNIME.

Objetivos de aprendizagem

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

- Conheçam a plataforma de análise de dados KNIME;
- Experimentem tarefas de exploração e de preparação de dados;

Enunciado

Descarregue o *dataset* disponível na plataforma de *e-learning* da Universidade do Minho, com dados sobre o fabrico e degustação de barras de chocolate.

Realize as tarefas seguintes:

- T1. Carregar o dataset «chocolat_bars» que apresenta dados sobre reviews de barras de chocolate;
- T2. Criar plots para visualização dos dados que respondam às seguintes perguntas:
 - a) Quais são os 3 ingredientes mais usados em conjunto? E de forma individual?
 - b) Quais são os 5 países com maior produção de grãos de cacau (bean_origin)?
 - c) Qual é o país (company_location) com maior média de avaliações (rating)?
 - d) Qual é o conjunto de ingredientes (ingredients) com maior média de avaliações (rating)?
- T3. Aplicar nodos de agregação de dados:
 - a) Por origem do grão de cacau (bean_origin), obter a quantidade de localizações (países) diferentes de empresas que o produzem (company_location);
 - b) Obter o rating que aparece com maior frequência;
 - c) Por ano de *review*, a média de *rating* obtido;
 - d) Por ano de review, obter a quantidade de reviews feitas;
 - e) Por barra de chocolate (*bar_name*), obter o *rating* médio para as barras com 100% de cacau (*cocoa_percent*)?
 - f) Por barra de chocolate (bar_name), obter o rating máximo obtido.
- T4. Aplicar nodos para tratamento de dados:
 - a) Excluir todas as colunas do tipo Integer,
 - b) Tratar valores em falta;
 - c) Remover registos duplicados;
 - d) Criar 3 bins de igual frequência para a feature "cocoa_percent";
 - e) Transformar o *rating* em 2 classes categóricas: se superior a 2,5 então é "good" senão é "bad".
- T5. Utilizar técnicas baseadas em árvores de decisão ("decision trees") para criar modelos de previsão:
 - a) Fazer a partição de dados (70/30) de forma aleatória;
 - b) Fazer o treino do modelo usando árvores de decisão;
 - c) Avaliar o resultado.

Descrição do *dataset* CHOCOLATE_BARS

ATRIBUTO	DESCRIÇÃO
id	Identificador da <i>review</i>
manufacturer	Nome da empresa produtora da barra de chocolate
company_location	Localização da empresa produtora
year_reviewed	Ano da <i>review</i>
bean_origin	País de origem dos grãos de cacau
bar_name	Nome da barra de chocolate
cocoa_percent	Percentagem de cacau na barra de chocolate
num_ingredients	Número de ingredientes na barra de chocolate
ingredients	Ingredientes usados para produzir a barra de chocolate
	B = Beans (grão de cacau), S = Sugar (açúcar)
	S* = Sweetener other than sugar or beet sugar (adoçante exceto açúcar ou açúcar de beterraba)
	C = Cocoa Butter (manteiga de cacau), V = Vanilla (baunilha)
	L = Lecithin (lecitina), Sa = Salt (sal)
review	Resumo das características mais memoráveis da barra de chocolate
rating	Avaliação da barra de chocolate

Mais detalhes sobre estes dados podem ser encontrados nesta ligação: flavorsofcacao.com/chocolate_bars