Data Mining e Big Data
Licenciatura em Informática

3º ano - 2º Semestre

Ano letivo 2024/2025

Universidade da Maia

Objetivo

Usar a metodologia CRISP-DM e descrever como se procederia em cada uma das fases da metodologia perante um problema de data mining usando o Rapidminer. Este projeto será trabalho individual.

Entrega

A documentação de cada projecto resume-se a um relatório PDF máximo 15 páginas (Letra Arial Tamanho 12 esp 1.5), e uma apresentação em Power Point que será realizada dia 12 de junho. Deverão também entregar os ficheiros Rapidminer.

Cada apresentação terá no máximo 15 minutos + discussão por grupo.
Grupos de 2 alunos máximo.

Estudo

Sugere-se a utilização do dataset *Dermatology* .

1) Na primeira fase da metodologia (*Business understanding*) deve ser apresentada uma breve descrição e contextualização do problema para se justificar a aplicação da metodologia CRISPDM.

2) Na segunda fase da metodologia (*Data understanding*) é necessário proceder à caracterização do problema (*dataset*) escolhido. Nomeadamente, esta caracterização deve conter:

   a) Qual o atributo objectivo (*label*) para o qual são geradas previsões, e as classes (diferentes valores para esse atributo), bem como o número de instâncias para cada valor do atributo label.

   b) Número de atributos e sua caracterização (tipos de dados, número de instâncias para cada valor do atributo, etc).

3) Na fase de Preparação de Dados recorrer às técnicas de pré-processamento de dados que são necessárias para proceder à aplicação da técnica de mineração de dados.

4) Na fase da Modelação: aplicar a técnica de classificação e o algoritmo C4.5 (*decision tree* no *Rapidminer*) variando o critério para dividir os dados: **gain_ratio**, **information_gain**, **gini_index** e **accuracy** e o resto dos parâmetros default.

5) Na fase de Avaliação analisar as performances obtidas de cada uma das árvores (optar pelo operador *Cross Validation com k folds 10* para a validação). Analisar os valores de *Accuracy*, *Classification error* e *Root Mean Squared Error*. Escolher e justificar qual a árvore escolhida e interpretar os resultados. Retirar conclusões a partir da árvore escolhida.

Detalhes da Base de Dados

O diagnóstico das doenças *erythemato-squamous* é um problema real na dermatologia. Todas as doenças compartilham aspectos clínicos com poucas diferenças. As doenças descritas nesta base de dados são *psoriasis*, *seboreic dermatitis*, *lichen planus*, *pityriasis rosea*, *cronic dermatitis* e *pityriasis rubra pilaris.* Normalmente é necessário realizar biópsia para o diagnóstico, mas infelizmente, muitas partilham características histopatológicas (tecidos lesionados). Outra dificuldade associada a este diagnóstico é que algumas das doenças partilham os mesmos sintomas em estágios iniciais e diferentes em estágios mais avançados. Os pacientes foram inicialmente avaliados clinicamente com 12 atributos (1: erythema, 2: scaling, 3: definite borders, 4: itching, 5: koebner phenomenon, 6: polygonal papules, 7: follicular papules, 8: oral mucosal involvement, 9: knee and elbow involvement, 10: scalp involvement, 11: family history, (0 or 1) e 34: Age (linear)). Posteriormente, foram colhidas amostras de pele para avaliação de 22 atributos histopatológicas e analisadas usando um microscópio (12: melanin incontinence, 13: eosinophils in the infiltrate, 14: PNL infiltrate, 15: fibrosis of the papillary dermis, 16: exocytosis, 17: acanthosis, 18: hyperkeratosis, 19: parakeratosis, 20: clubbing of the rete ridges, 21: elongation of the rete ridges, 22: thinning of the suprapapillary epidermis, 23: spongiform pustule, 24: munro microabcess, 25: focal hypergranulosis, 26: disappearance of the granular layer, 27: vacuolisation and damage of

basal layer, 28: spongiosis, 29: saw-tooth appearance of retes, 30: follicular horn plug, 31: perifollicular parakeratosis, 32: inflammatory monoluclear inflitrate, 33: band-like infiltrate).

Todos os atributos contém valores entre 0-3. O valor 0 indica que o recurso não estava presente, 3 indica a maior quantidade possível e 1, 2 indicam os valores intermediários relativos.

Excepto o atributo referente ao histórico familiar (atributo 11) que apresenta o valor 1 no caso de se verificar algum familiar que tenha sido diagnosticado com alguma das doenças apresentadas; e valor 0 caso se verifique o caso contrário. E o atributo referente à idade, que representa a idade do doente.

Os números de identificação e os nomes dos pacientes foram removidos da base de dados. Existem 8 instâncias que contém valores desconhecidos no atributo Idade (atributo 34 Age) representados por '?'.

A base de dados apresenta um total de 366 instâncias e 34 atributos (+ 1 atributo classe). Atributos disponíveis:

- atributo 1: erythema (valores possíveis: 0,1,2,3);
- atributo 2: scaling (valores possíveis: 0,1,2,3);
- atributo 3: definite borders (valores possíveis: 0,1,2,3);
- atributo 4: itching (valores possíveis: 0,1,2,3);
- atributo 5: koebner phenomenon (valores possíveis: 0,1,2,3);
- atributo 6: polygonal papules (valores possíveis: 0,1,2,3);
- atributo 7: follicular papules (valores possíveis: 0,1,2,3);
- atributo 8: oral mucosal involvement (valores possíveis: 0,1,2,3);
- atributo 9: knee and elbow involvement (valores possíveis: 0,1,2,3);
- atributo 10: scalp involvement (valores possíveis: 0,1,2,3);
- atributo 11: family history, (valores possíveis: 0, 1);

- atributo 12: melanin incontinence (valores possíveis: 0,1,2,3);
- atributo 13: eosinophils in the infiltrate (valores possíveis: 0,1,2,3);
- atributo 14: PNL infiltrate (valores possíveis: 0,1,2,3);
- atributo 15: fibrosis of the papillary dermis (valores possíveis: 0,1,2,3);
- atributo 16: exocytosis (valores possíveis: 0,1,2,3);
- atributo 17: acanthosis (valores possíveis: 0,1,2,3);
- atributo 18: hyperkeratosis (valores possíveis: 0,1,2,3);
- atributo 19: parakeratosis (valores possíveis: 0,1,2,3);
- atributo 20: clubbing of the rete ridges (valores possíveis: 0,1,2,3);
- atributo 21: elongation of the rete ridges (valores possíveis: 0,1,2,3);
- atributo 22: thinning of the suprapapillary epidermis (valores possíveis: 0,1,2,3);
- atributo 23: spongiform pustule (valores possíveis: 0,1,2,3);
- atributo 24: munro microabcess (valores possíveis: 0,1,2,3);
- atributo 25: focal hypergranulosis (valores possíveis: 0,1,2,3);
- atributo 26: disappearance of the granular layer (valores possíveis: 0,1,2,3);
- atributo 27: vacuolisation and damage of basal layer (valores possíveis: 0,1,2,3);
- atributo 28: spongiosis (valores possíveis: 0,1,2,3);
- atributo 29: saw-tooth appearance of retes (valores possíveis: 0,1,2,3);
- atributo 30: follicular horn plug (valores possíveis: 0,1,2,3);
- atributo 31: perifollicular parakeratosis (valores possíveis: 0,1,2,3);
- atributo 32: inflammatory monoluclear inflitrate (valores possíveis: 0,1,2,3);
- atributo 33: band-like infiltrate (valores possíveis: 0,1,2,3);
- atributo 34: Age (valores possíveis: inteiros positivos).

Distribuição de Classes (Atributo especial *label*):

- Valor 1 corresponde a 'psoriasis';
- Valor 2 corresponde a 'seboreic dermatitis';
- Valor 3 corresponde a 'lichen planus';

- Valor 4 corresponde a 'pityriasis rosea';
- Valor 5 corresponde a 'cronic dermatitis';
- Valor 6 corresponde a 'pityriasis rubra pilaris'.