FE06

Universidade do Minho
Escola de Engenharia

Curso: Mestrado em Engenharia Biomédica
U.C.: Aprendizagem e Extração do Conhecimento

Plano de Aula 06	
Docente:	Hugo Peixoto | José Machado
Tema:	RapidMiner: Classificação
Ano Letivo:	2023-2024 – 1º Semestre
Duração da aula:	2 horas

1. Introdução

O Data Set usado neste exercício é sobre doenças cardíacas disponível no ficheiro heart-c.csv, obtido através do repositório da UCI. Este Data Set descreve fatores de risco para doenças cardíacas. O atributo ***num*** representa o atributo da classe (binominal):

class <50 - nenhuma doença
class> 50_1 - aumento do nível de doença cardíaca.

O principal objetivo deste exercício é prever doenças cardíacas a partir de outros atributos no Data Set. Obviamente, trata-se de um problema de classificação. O software a ser usado é o RapidMiner. A descrição deste exercício é gradual. Portanto, espera-se que possa entender melhor os vários aspetos e questões envolvidos no processo de Extração de Conhecimento.

2. Data Understanding

O primeiro passo para abordar o problema é familiarizar-se com os dados. Responder às seguintes perguntas ajudará a entender melhor os dados.
Aceda ao url do Data Set e perceba todos os atributos envolvidos. De seguida carregue o Data Set no RapidMiner.

[1] Para cada atributo, encontre as informações inerentes ao processo de Data Understanding, recorrendo às "Estatísticas" e "Visualizações" do RapidMiner:

[a] O tipo de atributo, Percentagem de valores ausentes nos dados, máximo, mínimo, média, moda e desvio padrão (quando aplicável). Poderá apresentar estes resultados em formato de tabela.

[b] Existem instâncias que tenham um valor para um determinado atributo que nenhuma outra instância tem, i.e. registos únicos?

[2] Estude os diferentes histogramas (atributos numéricos) e descreva informalmente como alguns atributos parecem influenciar o risco de doença cardíaca. Poderá documentar as observações através da captura dos gráficos.

[3] Utilizando o gráfico de dispersão "Scatter", avalie as seguintes questões:
[a] Caso considere possível indique se o atributo *thalach* parace estar mais/menos associado a doenças cardíacas? E o atributo *ca*?

[4] Observando o gráfico thalach(X) / oldpeak (y) Investigue uma possível associação desses atributos com o atributo class, ou seja, tente identificar possíveis áreas "densas" de doenças cardíacas (se existirem).

3. Data Processing

A segunda etapa diz respeito ao processamento dos dados de modo que os dados transformados estejam numa forma mais adequada para os algoritmos de data mining. Todos as alíneas devem partir do ficheiro original e ser incluídas num subprocesso à semelhança do executado em aulas anteriores. O resultado deste processo será no fim alimento aos algoritmos de ML para avaliação dos modelos desenvolvidos. Para cada uma das alíneas seguintes documente com printscreens.

[1] Lidar com valores ausentes

[a] Os registos nulos não deverão ser eliminados e é aconselhável atribuir valores onde faltam dados, usando um método adequado. Utilize um operador para substituir os valores ausentes pela média do atributo, se o atributo for numérico. Caso seja nominal deverá substituir pela moda. Este modelo, deve ser guardado para ser usado posteriormente na avaliação. Guarde o processo no RapidMiner com o nome fe06-p1.

[2] Avaliação do peso dos Atributos

[a] Investigue a possibilidade de usar o operador "*Weight by Information Gain*" para selecionar um subconjunto de atributos com boa capacidade de previsão.

- Deve considerar um valor de peso superior a 0,4 (Guarde o processo no RapidMiner com o nome fe06-p2)

[b] Compare o subconjunto de atributos escolhido, com os resultados obtidos na questão [2] da Parte I.

[3] Eliminar outliers

Investigue a possibilidade de usar o operador "Detect Outlier" para detectar outliers. Ao efetuar a aplicação do filtro encontrará instâncias classificadas como *outliers* = "true". Utilizando o operador "Filter Examples" para criar um novo processo com o Data Set sem *outliers*. (Guarde o processo no RapidMiner com o nome fe06-p3)

[4] Normalização

Investigue a possibilidade de usar o operador "Normalize" para normalizar os valores presentes no Data Set. (Guarde o processo no RapidMiner com o nome fe06-p4)

[5] Processo Extra

Utilize um processo que combine os passos 2, 3 e 4 e guarde esse processo. (Guarde o processo no RapidMiner com o nome fe06-p5).

4. Modeling

O terceiro passo é usar algoritmos de classificação disponíveis no RapidMiner para descobrir padrões ocultos nos dados. Deve repetir as etapas descritas abaixo para cada um dos processos criados durante a parte 2, além de usar também o Data Set original.

[1] Comece com o classificador "Decision Trees".

[a] O que pode concluir? Compare as suas conclusões (atributos importantes) com as conclusões que obteve na Parte I questão 2.

[b] Compare a precisão do classificador obtida usando o operador "Split Data" e o operador "*Cross Validation*". Como explica esta diferença (se existir)?

[c] Crie um classificador com e sem pruning. Qual se comporta melhor? Justifique a sua resposta.

[2] Explore outros 2 algoritmos de classificação, usados previamente nas aulas, e vá guardando os resultados. Utilize os Métodos de "Split Data" e "Cross Validation" para todos.

5. Evaluation

Nas etapas [3] e [4] construiu vários modelos. Por fim, é necessário comparar os diferentes modelos e apresentar as suas conclusões.

[1] Escolha algumas medidas de desempenho e justifique a sua escolha.

[2] Resuma numa tabela as medidas de desempenho para cada classificador e cada Data Set.

[3] O que pode concluir?