Mineração de Dados e Descoberta de Conhecimento

Profs. Heitor S. Lopes e Thiago H. Silva (UTFPR, 2022) - Exercício #3

A) Árvores de decisão

1. **Objetivo**: explorar a classificação de dados com árvores de decisão usando o software Orange.
2. **Procedimentos**:
   a. O *dataset* *Soybean* se refere ao diagnóstico de 19 doenças comuns da soja. Ele tem 35 atributos e 683 instâncias. Faça o pré-processamento necessário para *upload* no Orange.
   b. Utilizando as ferramentas de visualização de dados, o que é possível preliminarmente inferir preliminarmente sobre os atributos deste dataset?
   c. Selecione a coluna "*class*" como o alvo da classificação, sendo as demais colunas os atributos previsores. Use validação cruzada estratificada de 5-*folds* para o treinamento de uma Árvore de Decisão com os parâmetros *default*. Anote o tamanho da árvore obtida (número total de nós, profundidade e número de nós-folhas) e as medidas de qualidade (acurácia, *precision*, *recall* e F1 *score*). Justifique qual a medida de qualidade adequada para este caso.
   d. Mostre a matriz de confusão gerada pelo treinamento/teste da árvore de decisão. Identifique nesta árvore quais foram as classes que tiveram 100% e 0% de acerto, respectivamente. Justifique este comportamento (em especial para as classes com 0% de acerto).

B) Regras de classificação

3. **Objetivo**: explorar a classificação de dados com árvores de decisão e regras de classificação com o software Orange.
4. **Procedimentos**:
   a. Utilize o *Contraceptive Method Choice Dataset*, disponível no *Machine Learning Repository*. O objetivo aqui é predizer o método contraceptivo utilizado por mulheres da Indonésia, com base em indicadores demográficos e sócio-econômicos. O *dataset* tem 1473 instâncias e 9 atributos previsores, e 3 classes desbalanceadas (*No-use, Long-term e Short-term*).
   b. Utilizando as ferramentas de visualização de dados, o que é possível preliminarmente inferir preliminarmente sobre os atributos deste dataset?

c. Utilizando a ferramenta de discretização, faça um pré-processamento dos atributos numéricos *wifes_age* e *number_chd_born*, discretizando os dados em faixas com **igual frequência,** respectivamente em: {jovem, adulta, madura, meia-idade} e {1, 2, 3_4, 5+}.
d. Utilize os algoritmos *baseline* (ZeroRule e OneRule) para estabelecer um referencial de comparação acerca da qualidade da classificação. Anote as métricas de qualidade (acurácia, *precision*, *recall* e F1 *score*).
e. Utilize o CN2 para gerar regras de classificação que sejam **compreensíveis** e **interessantes** (mas, **não óbvias**). Com base nestas regras, contextualize e esclareça o perfil (sócio-econômico e cultural) das usuárias de métodos contraceptivos *short_term* e principalmente *long_term* na Indonésia.
f. Compare a qualidade preditiva dos três métodos (ZeroRule, OneRule e CN2).
g. Gere uma árvore de decisão com uma profundidade que permita a sua compreensão. Compare o resultado aqui obtido com as regras de classificação anteriormente obtidas e discuta os pontos positivos e negativos das duas abordagens (árvores X regras) em termos de qualidade preditiva e compreensibilidade.