Trabalho Prático de Mineração de Dados

Professor: Manoel Limeira

O trabalho poderá ser realizado em dupla e deverá contemplar quatro tarefas da Mineração de Dados:
(A) Classificação e
(B) Regressão
(C) Regras de Associação
(D) Clusterização

Ferramenta sugerida:

- Orange disponível em: https://orange.biolab.si

Os exemplos das aulas práticas serão demonstrados na ferramenta Orange.

1ª Fase: seleção e pré processamento da base de dados
Escolher uma base de dados disponível em algum repositório, como por exemplo:
UCI (University of California, Irvine) disponível em: http://www.ics.uci.edu/~
Kaggle: https://www.kaggle.com/

Atributos, instâncias e contexto das bases devem ser devidamente descritos.
Cada dupla deverá escolher uma base diferente. Pode ser a mesma base para as três tarefas. Para garantir que não haverá bases duplicadas, cada dupla deve informa o nome da base(s) escolhida(s).

2ª Fase: execução das tarefas de mineração
(A) Classificação: executar pelo menos 5 algoritmos de classificação, fazendo uma comparação entre os resultados obtidos para a base adotada. Para cada algoritmo, verificar se é possível fazer variações dos parâmetros de entrada.

(B) Regressão: executar pelo menos 5 algoritmos de regressão, fazendo uma comparação entre os resultados obtidos para a base adotada. Para cada algoritmo, verificar se é possível fazer variações dos parâmetros de entrada.

(C) Associação: executar algoritmo de extração de regras de associação.
Explorar as principais (entre 5 e 10) medidas de interesse disponíveis, com diferentes valores mínimos, e analisar os resultados.

(C) Clusterização: executar dois algoritmos de clusterização. Caso escolham uma base que já tenha um atributo classe, não deixar de desconsiderar esse atributo na clusterização. Caso contrário, a clusterização será influenciada por esse atributo e o exercício não fará sentido. Verificar se a clusterização coincidiu com a divisão anterior de classes e variar os parâmetros de entrada dos algoritmos selecionados. Analisar os resultados com base nos centroides dos clusters.

IMPORTANTE:
Data de entrega do relatório com os resultados obtidos em formato PDF: 25/02/2020.