P.PORTO ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO	**LEI – Licenciatura em Engenharia Informática** IA – Inteligência Artificial 2º Semestre - Docente: DCarneiro Ficha Prática 2

Tema: Introdução ao Machine Learning e à preparação de dados.

Objetivos: Familiarização com a aplicação Weka e com o ecrã de pré-processamento. Aplicação das principais tarefas de pré-processamento.

Exercício 1

Considere o dataset Titanic, disponibilizado no Moodle e cujas variáveis estão descritas em https://www.kaggle.com/c/titanic/data. Este dataset detalha alguma informação sobre os passageiros do Titanic, incluindo os que sobreviveram ou não. O objetivo deste exercício é preparar o dataset para a tarefa de prever se um dado passageiro sobrevive ou não.

Implemente as seguintes tarefas de preparação dos dados:

- Note que as variáveis Survived, Pclass apesar de serem categorias, são representadas numericamente. Transforme as variáveis em categorias, aplicando o filtro não supervisionado NumericToNominal.
- Para facilitar a visualização dos dados, aplique um filtro de discretização supervisionado na variável Age.
    - Interprete os resultados da operação anterior
    - Anule o filtro anterior e aplique agora um filtro de discretização não supervisionado, com 3 bins.
    - Interprete os resultados. Ganhou-se algo em termos de representação/interpretação/visualização dos dados?
- Note que a variável Age tem 20% de dados em falta. Decida, de forma fundamentada, como tratar os dados. Algumas das opções possíveis são:
    - Remover as instâncias com dados em falta na variável Age: unsupervised.instance.RemoveWithValues
    - Preencher os valores em falta com o valor médio de idade (Impute): unsupervised.attribute.ReplaceMissingValues
- As variáveis Name, Ticket, PassengerID e Cabin não terão, à partida, qualquer influência no problema a tratar. Remova-as.
- Que conclusões preliminares se podem retirar de uma análise visual dos dados?
- Guarde o dataset com um novo nome, para o poder utilizar mais tarde.

Exercício 2

Considere o dataset desenvolvido no Exercício 1. Abra o ecrã Visualize. Selecione visualizações que tentem responder às seguintes questões:

- Há alguma relação entre a idade, o preço do bilhete e a probabilidade de sobrevivência?

- Qual a relação entre a idade, o género, e a probabilidade de sobrevivência?
- Qual a relação entre a classe do bilhete, o género, e a probabilidade de sobreviver? Utilize a opção Jitter para facilitar a visualização dos dados.

Exercício 3

Considere o dataset desenvolvido no Exercício 1.

- Através do ecrã Classify, selecione o algoritmo classifiers.functions.MultiLayerPerceptron. Selecione a variável de classe apropriada para o problema e treine o modelo com as seguintes alterações às configurações:
    - GUI: True
    - trainingTime: 5000
- Interprete os resultados obtidos.
- Guarde o modelo para poder usá-lo mais tarde.

Exercício 4

Repita o exercício anterior mas agora utilizando um modelo classifiers.functions.J48.

- Visualize a árvore resultante e interprete os resultados obtidos.
- Guarde o modelo para poder usá-lo mais tarde.