

## PROVA 2 - INTELIGÊNCIA ARTIFICIAL

Elabore um único relatório com todos os procedimentos e resultados obtidos em cada questão, argumentando os resultados. Só será aceito o relatório com as soluções das questões. As questões podem ser implementadas utilizando qualquer linguagem de programação ou o *Weka*.

### Classificação

1. Utilizando a base disponível em: <https://www.kaggle.com/c/titanic/data> elabore uma solução utilizando dois algoritmos de aprendizagem de máquina do seu conhecimento para classificar se o passageiro tem o não chance de sobreviver. Os resultados dessa questão deverão ser descritos detalhadamente no relatório através de três pontos principais:

- Análise da base de dados: identificar instâncias com atributos incompletos, gerar matriz de correlação, identificar a presença de *outliers* e verificar se as classes estão balanceadas. O balanceamento deverá ser ilustrado por meio de gráficos (e.g. histograma);
- Justifique a escolha dos dois algoritmos de aprendizagem de máquina utilizados e discuta os resultados obtidos em ambos. Explique porque você acredita que os algoritmos escolhidos são mais apropriados para o problema;
- Analise os resultados considerando matriz de confusão, especificidade, sensibilidade, medida f1 e acurácia. Descreva detalhadamente os resultados obtidos por cada métrica, justificando a diferença entre eles.

Dicas:

- A partir da análise da base de dados, para bons resultados possivelmente será necessário pré-processar os dados. Para identificar os *outliers*, recomenda-se ilustrá-los graficamente (e.g. boxplot).
- Divida (aleatoriamente) o conjunto em treinamento e teste: 80% e 20%, respectivamente.
- Para analisar os resultados de cada métrica, identifique os falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

**Bônus:** Altere dois parâmetros de cada algoritmo de aprendizagem de máquina utilizados na questão e discuta os resultados obtidos. Exemplo: alterar a quantidade de k vizinhos e a função de distância utilizada, alterar a função kernel do SVM, alterar a arquitetura da rede neural (e.x. camadas e função de ativação, alterar o otimizador e a taxa de aprendizado).

### Regressão

2. Utilizando as bases de treinamento e testes disponível no link: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>, apresente duas soluções de aprendizagem de máquina que consigam baixas taxas de erro médio (prove que esse é o menor valor que você pôde encontrar).

Dica:

- Teste com dois algoritmos de aprendizagem de máquina para identificar os que apresentam melhores resultados;

### Clusterização

3. Utilizando a base: <https://www.kaggle.com/thesiff/premierleague1819>. Execute os algoritmos de clusterização citados a seguir e compare os resultados.

○ Execute os algoritmos de agrupamento K-means e Hierárquico com os seguintes valores de K: 2, 5, 10 e 100. Compare os agrupamentos resultantes dos 2 algoritmos.

- Escolha um número fixo de K e altere o parâmetro do K-Means referente ao número máximo de iterações: 1, 10 e 100 e o parâmetro de Linkage do Hierárquico, quais diferenças puderam ser observadas?
- Faça uma comparação entre os 2 algoritmos, qual você acha que teve o melhor desempenho e por quê?

**Bônus:** normalizar os dados e executar novamente os algoritmos para analisar os novos resultados gerados. Em seguida, fazer os gráficos dos agrupamentos dos dados brutos vs dados normalizados, e um boxplot para mostrar a dispersão destes dados.