# Primeira prova de Ciência de Dados - 24/09/2024

Todas os códigos devem estar em um único arquivo `.R`. Além do código em `.R`, os gráficos solicitados também devem ser anexados à tarefa.

**Questão 1.** O conjunto `baleias.txt` apresenta dados sobre baleias. Cada linha do arquivo contém informações sobre uma baleia, com as seguintes variáveis: `especie` (espécie da baleia), `comprimento` (comprimento da baleia), `peso` (peso da baleia), `profundidade_maxima` (profundidade máxima de mergulho) e `volume_cranio` (volume do cranio da baleia). Faça uma análise gráfica dos dados, para caracterizar as espécies de baleias. Anexe os gráficos à tarefa e deixe todos as respostas dissertativas como comentários no código.

**Questão 2.** Queremos analisar os fatores que influenciam a saída de clientes (churn) de uma instituição financeira. Para isso, temos o conjunto de dados do arquivo `churn.txt`. O conjunto de dados inclui informações sobre a pontuação de crédito do cliente (CreditScore), sua localização geográfica (Geography), gênero (Gender) e sobrenome (Surname). Também temos dados sobre a idade do cliente (Age) e o tempo que ele está com a empresa (Tenure). Informações financeiras são representadas pelo saldo da conta do cliente (Balance), o número de produtos que ele possui com o banco (NumOfProducts), se possui cartão de crédito (HasCrCard, onde 1 = Sim, 0 = Não) e seu salário estimado (EstimatedSalary). O nível de engajamento do cliente é indicado pela variável IsActiveMember (1 = Sim, 0 = Não). Há também uma variável chamada RowNumber que enumera as linhas do conjunto e uma variável que apresenta o número de identificação do cliente (CustomerId). Nossa variável alvo é Exited, que indica se o cliente saiu da instituição (1 = Sim, 0 = Não). Esta é a variável que queremos entender e potencialmente prever com base nas outras informações disponíveis.

(a) Importe o arquivo para o R, entenda a estrutura do conjunto e faça mudanças necessárias em suas variáveis.

(b) Crie um modelo de árvore de decisão para prever a variável `Exited`. O modelo deve ser construído com um conjunto de treinamento que contenha 75% dos dados. Avalie a acurácia do modelo e construa a matriz de confusão. Analise a acurácia e a matriz e comente os resultados obtidos. Anexe à prova o gráfico da árvore de decisão.

(c) A partir do conjunto que contém todos os dados, crie uma data.frame que conterá as informações para os clientes de cada país. Exemplo, crie um data.frame apenas com os clientes da França. Para cada um desses conjuntos, crie um modelo de árvore de decisão para prever a variável `Exited` (75% para treinamento). Avalie a acurácia do modelo e construa a matriz de confusão. Analise a acurácia e a matriz e comente os resultados obtidos. Há diferença na previsão entre os países? Há diferença nas previsões encontradas em (c) e na previsão encontrada em (b)? Anexe à prova os gráficos das árvores de decisão.