

**Disciplina:** Aprendizagem de Máquina  
**Período:** 2022.1  
**Professor:** César Lincoln Cavalcante Mattos

### Lista 3 - KNN e árvores de decisão

#### Instruções

- Com exceção dos casos explicitamente indicados, os algoritmos e modelos devem ser implementados do início em qualquer linguagem de programação (Python, R, Octave...).
- Pacotes auxiliares (sklearn, matplotlib, etc) podem ser usados somente para facilitar a manipulação dos dados e criar gráficos.
- A entrega da solução pode ser feita via pdf ou Jupyter notebook pelo SIGAA.

### Observações

- **Graduação:** Sem mudanças.
- **Pós-graduação:** Considere a distância Euclidiana e a distância de Mahalanobis no KNN.

### Questão 1

Considere o conjunto de dados disponível em **kc2.csv**, organizado em 22 colunas, sendo as 21 primeiras colunas os atributos e a última coluna a saída. Os 21 atributos são referentes à caracterização de códigos-fontes para processamento de dados na NASA. A saída é a indicação de ausência (0) ou existência (1) de defeitos. Maiores detalhes sobre os dados podem ser conferidos em <https://www.openml.org/d/1063>.

- Considerando uma validação cruzada em 10 *folds*, avalie modelos de classificação binária nos dados em questão. Para tanto, use as abordagens abaixo:
  - **KNN** (escolha  $k = 1$  e  $k = 5$ , distância Euclidiana (e Mahalanobis, para a pós-graduação));
  - **Árvore de decisão** (você pode usar uma implementação já existente com índices de impureza de gini e entropia).
- Para cada modelo criado, reporte valor médio e desvio padrão das métricas de **acurácia**, **revocação**, **precisão** e **F1-score**.