**Lista 1 - Tarefa de classificação**

**Objetivo:** Aprender a aplicar técnicas de aprendizado de máquina para a tarefa de classificação, classificando as máquinas que possuem falhas e as que não possuem falhas.

**Conjunto de dados:** utilizar o conjunto de dados “predictive\_maintenance.csv”, disponível em [Machine Predictive Maintenance Classification | Kaggle](https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification).

**Instruções:**

1. **Faça o download do conjunto de dados citado anteriormente;**
2. **Responda às questões na sequência, fazendo as devidas análises demandadas;**
3. **Salvar o código com as respostas (devidamente identificadas no arquivo) utilizando o seguinte nome de arquivo: “nome\_do\_aluno\_CLASSIFICAÇÃO.ipynb”;**
4. **Postar o arquivo na plataforma Moodle, dentro da pasta de atividade, até o dia 17 de Abril de 2022.**
5. Importe as bibliotecas necessárias (*pandas*, *matplotlib.pyplot*, *sklearn*, etc.).
6. Considere a primeira coluna como o índice do *DataFrame* e desconsidere a coluna “Product ID” e também a coluna “Failure Type”.
   1. Veja no *site* do *Kaggle* indicado o significado de cada um dos atributos deste conjunto.
7. Faça o tratamento e exploração dos dados da maneira que achar mais conveniente (transformação, normalização, exclusão de dados faltantes, imputação de dados, binarização, etc).
8. Prepare o modelo para a validação *holdout*, considerando 85% do conjunto para treinamento e parâmetro *random­\_state* = 5.
   1. Observação: verifique o balanceamento de classes para ver se não há uma discrepância muito alta. Caso a classe mais favorecida tenha mais do que 10 vezes o número de observações da classe menos favorecida, realize o balanceamento pela técnica SMOTE (*Synthetic Minority Oversampling Technique*) para que a classe menos favorecida tenha 30% de instâncias em relação à classe mais favorecida.
9. Use os métodos de árvores de decisão, floresta randômica e *k*-vizinhos mais próximos para classificar o conjunto.
   1. Para a árvore de decisão, mostre a árvore de decisão criada com uma profundidade de quatro níveis (*max\_depth*), destacando qual é o atributo do nó raiz da árvore.
   2. A partir dos resultados obtidos pelos modelos, responda:
      1. Qual obteve a maior acurácia?
      2. Qual modelo obteve a maior sensibilidade (*recall*) e precisão?
      3. Qual retornou a melhor média harmônica entre *recall* e precisão?
      4. Mostre quais foram as frequências de Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN).
   3. Caso tenha sido realizado o balanceamento, aplique as três técnicas também para o conjunto original (sem o *SMOTE*), e compare as métricas entre os modelos criados.