Universidade Federal do Ceará.
Campus de Quixadá
Disciplina: Aprendizado de Máquina
Professor: Regis Pires Magalhães **Data: ____/____/_____**
Aluno: _____ **Matrícula: _____**

Avaliação Parcial 1

Importante:

- Acesse o endereço: `http://albertao.quixada.ufc.br/ml` e copie os arquivos `mlap1.zip` e `mldocs.zip` para seu computador.
- Desconecte o cabo de rede do computador.
- Somente é permitida consulta à documentação contida na pasta `mldocs`.
- Não é permitido o uso de rede, pendrive ou qualquer outro meio de armazenamento externo de dados.
- Ao concluir a avaliação, compacte somente a sua pasta `mlap1` com *dataset* e resolução, e altere o nome do arquivo compactado para conter a matricula e o nome do aluno (`<matr>-<nome>-mlap1`). Somente depois disso, chame o professor para entregar sua resolução, que deverá ser copiada para o pendrive do professor.
- Sempre que possível, use uma semente (*seed*) ou *random_state* com o valor 42.
- Crie um Jupyter Notebook para responder as questões a seguir.

1. O *dataset breast cancer wisconsin_ap1.csv* contém valores faltando e rótulo (*label*) 'M' para classe Maligno e 'B' para classe Benigno. Faça atribuição da média da coluna para valores faltantes. Adeque o *label* para que possa ser devidamente usado por diversos algoritmos de aprendizado de máquina. (2 pontos)

2. Responda os itens a seguir (1 ponto):
a) Que atributo (*feature*) possui maior valor absoluto de correlação com o *label*?
b) Que atributos (*features*) possuem maior valor absoluto de correlação entre si?
c) Qual *feature* do *dataset* mais se assemelha a uma distribuição normal (gaussiana)? Explique textualmente sua resposta e, se possível, mostre algum gráfico que ajude na sua explicação textual.

3. Use 75% do dados para treino e 25% para teste sem validação cruzada, mas **com estratificação** sobre os rótulos do dataset *breast cancer wisconsin_ok.csv* (2 pontos):
a) Faça *Standardization* dos dados.
b) Criar e treinar modelos preditivos para os dados com e sem *Standardization,* usando os seguintes algoritmos (`sklearn.linear_model...`): *Perceptron*, *Stocastic Gradient Descent (SGD)* e *Logistic Regression*.

4. Prove que os dados dos seus conjuntos de dados estão devidamente **estratificados** (1 ponto).

5. Implemente uma função para calcular a acurácia, usando somente Python e/ou NumPy, mas sem usar a biblioteca *scikit-learn*. A função deve receber como parâmetros *y_real* (array com rótulos reais) e *y_predito* (array com rótulos preditos). (1,5 pontos)

6. Calcular a acurácia (preferencialmente usando sua implementação) e mostrar o valor da acurácia para cada algoritmo sobre (1 ponto):
   a) o conjunto de treino não estandardizado.
   b) o conjunto de treino estandardizado.
   c) o conjunto de teste não estandardizado.
   d) o conjunto de teste estandardizado

7. Com relação aos resultados da questão anterior responda (1,5 ponto):
   a) Qual algoritmo apresentou a pior acurácia?
   b) Qual algoritmo apresentou a melhor acurácia?
   c) Qual conjunto de dados apresentou a pior acurácia?
   d) Qual conjunto de dados apresentou a melhor acurácia?
   e) Ocorreu *overfitting* em seus resultados? Explique sua resposta.