Universidade Federal de Alagoas

Instituto de Computação

Ciência de Dados

Professor: Bruno Pimentel

Aluno: _____

Lista de Exercícios 1

1. Qual a diferença entre *Big Data* e Ciência de Dados? (0,5 ponto)

2. De que forma Estatística, Mineração de Dados e Aprendizagem de Máquina interagem com Ciência de Dados? (1 ponto)

3. Mostre a importância do conhecimento de domínio para o cientista de dados. (0,5 ponto)

4. Crie um conjunto de dados com duas variáveis V1 e V2, tal que:

    a. Mediana de V1 < Média de V1 (0,5 ponto)

    b. Mediana de V2 > Média de V2 (0,5 ponto)

5. Baseando-se no conjunto de dados criado na questão 4, crie uma função em Python que:

    a. Mostra o histograma de cada variável; (1 ponto)

    b. Verifica se as variáveis seguem uma distribuição Normal (use teste de hipótese) (1 ponto)

6. Cite 2 técnicas para remoção de ruídos e, para cada uma, mostre uma vantagem e uma desvantagem. (1 ponto)

7. Qual é a importância de utilizar as seguintes abordagens de redução de dados no contexto de Ciência dos Dados?

    a. Redução de dimensionalidade (0,5 ponto)
    b. Redução de numerosidade (0,5 ponto)

8. De que forma pode-se detectar *overfitting* em um classificador? (0,5 ponto)

9. Em quais tipos de problemas é preferível utilizar *leave-one-ou*t a utilizar *K-fold cross-validation*? (0,5 ponto)

10. Crie um *script* em Python que avalie a diferença de desempenho do classificador K-NN e Naive Bayes para o conjunto de dados Iris (https://archive.ics.uci.edu/ml/datasets/iris). Use *F-measure* e *K-fold cross-validation.* (2 pontos)