UNIVERSIDADE FEDERAL DA PARAÍBA - UFPB
CENTRO DE INFORMÁTICA - CI
João Pessoa, __/10/2021

PROVA 02 - INTRODUÇÃO A INTELIGÊNCIA ARTIFICIAL

Escolha a questão 01 **OU** 02 (Utilize validação cruzada estratificada 10-fold e fixe o random_state) para responder, além da 03, em um notebook, enviando o link ou o código por e-mail.

Nome:
Matrícula:

1. Utilizando a base de dados de Churn Modelling (https://www.kaggle.com/shrutimechlearn/churn-modelling/version/1), elabore uma solução para identificar se o cliente deixou o banco ou se sua conta ainda está ativa. Lembre-se de comentar seu código no notebook detalhadamente, explicando cada passo.

   I. Faça o pré-processamento dos dados (limpeza, engenharia de variáveis, etc) e deixe os seus dados preparados para aplicar o modelo.

      OBS: Crie, pelo menos, UMA nova variável e explique o seu raciocínio.

   II. Faça uma breve análise exploratória dos dados, utilizando pelo menos dois gráficos.

   III. Escolha dois modelos de classificação para fazer a previsão de churn. Para cada algoritmo, faça ajustes em 2 hiperparâmetros.

   IV. Para avaliar os resultados, utilize e explique a matriz de confusão. Além disso, escolha uma outra métrica de sua preferência e o que o seu resultado significa.

2. No seguinte https://www.kaggle.com/hely333/eda-regression/data temos um conjunto de dados que contém o custo do tratamento de diversos pacientes. Como sabemos n variáveis podem afetar este valor, seja o diagnóstico do paciente, o local onde ele reside, ou o hospital ou clínica onde foi feito o tratamento, além de outros fatores. Sendo assim, vocês devem explorar os dados contidos nesta base e fazer uma análise de regressão a fim de prever o custo do tratamento dos pacientes.

   I. Realizem todo o pré processamento necessário para a utilização da base (verifiquem dados faltando, dados duplicados, etc).

II. Analisem os dados criteriosamente e tentem descobrir quais variáveis são mais relevantes para o resultado final.
III. Escolha ao menos dois modelos de regressão para a execução deste trabalho e compare seus resultados. Para cada algoritmo, faça ajustes em 2 hiperparâmetros.
IV. Utilizem o MSE ou Erro Quadrático Médio como uma métrica de avaliação do seu modelo.

OBS.: Lembrem-se de detalhar todos os passos e decisões tomadas durante a atividade, além de explicar cada gráfico utilizado e resultado obtido.

3. (TEM QUE) Utilize a seguinte base de dados: https://www.kaggle.com/akram24/mall-customers, que trata de informações de clientes de um shopping, que contém várias informações interessantes, como:

- ID;
- Sexo;
- Idade;
- Renda anual do cliente;
- Pontuação atribuída pelo shopping com base no comportamento do cliente e a natureza de seus gastos.

Diante disso:

I. Com seu conhecimento sobre os algoritmos de agrupamento, execute o K-means e Hierárquico.

II. Altere a quantidade de clusters para 3 valores de sua escolha.

III. Na execução do Hierárquico, varie 2 métodos do linkage; **OBS.: utilize os mesmos valores de clusters escolhidos na questão anterior.**

IV. Por fim, faça uma comparação entre os 2 resultados das execuções anteriores e adote uma medida de avaliação própria para clusterização.