EXERCÍCIOS - CLASSIFICAÇÃO

Para este exercício, será utilizado o dataset de **sobreviventes do titanic**. Os dados já encontram-se separados em arquivos de treino (train.csv) e teste (test.csv). Neste dataset, encontram-se informações como sexo, idade, classe socio-ecônomica, entre outras. Abaixo, você encontra o dicionário dos dados, contendo a descrição do que cada atributo e seus respectivos valores representam.

Atributo	**Descrição**	**Valores**
Survived	Indicador de sobrevivência do passageiro	0 = Não, 1 = Sim
Pclass	Classe do ticket	1 = 1ª classe, 2 = 2ª classe, 3 = 3ª classe
Sex	Sexo da pessoa	
Age	Idade em anos	
Sibsp	Quantidade de irmãos/ cônjuges a bordo do Titanic	

Parch	Quantidade de pais/filhos a bordo do Titanic	
Ticket	Número do ticket	
Fare	Tarifa do passageiro	
Cabin	Número da cabine do passageiro	
Embarked	Porto de embarcação	C = Cherbourg, Q = Queenstown, S = Southampton

Observações:

- Age: se a idade for menor que 1, o valor é fracional. Se a idade for uma estimativa, ela estará na forma xx.5.Age
- Sibsp: neste atributo, as relações familiares são definidas da seguinte forma..
    - Sibling = irmão, irmã, meio-irmão, meio-irmã;
    - Spouse = esposo, esposa.

- Parch: neste atributo, as relações familiares são definidas da seguinte forma..
    - Parent = mãe, pai.
    - Child = filha, filho, enteada, enteado;
- Algumas crianças embarcaram apenas com a suas babás logo, para esses casos, Parch = 0.

Utilizando o dataset fornecido, faça o que se pede:

1. Execute uma **análise exploratória** dos dados. Elabore gráficos, calcule estatísticas e obtenha inferências iniciais sobre os dados. Discorra sobre as inferências realizadas.
2. Realize o pré-processamento dos dados. Faça as limpezas e formatações que julgar necessárias para obter um conjunto de dados consistente. (**Dica:** você pode juntar os dois arquivos de dados em um único dataframe para facilitar a manipulação!).
3. Crie um classificador para predizer se um passageiro **sobreviveu ou não** a partir dos atributos presentes no dataset. Utilize os algoritmos KNN, Regressão Logística e Naive Bayes para criar os modelos. Crie um modelo para cada algoritmo.
4. Obtenha as métricas de avaliação de cada modelo criado (acurácia, kappa, F1, recall, precisão, falsos positivos, falsos negativos, quantidade de instâncias classificadas corretamente). Plote um mapa de calor exibindo a matriz de confusão de cada um dos modelos. O que você pode inferir dos modelos a partir das métricas obtidas? Explique suas respostas.
5. Divida os conjuntos de treino e teste usando **Kfold.split**. Realize os mesmos passos dos itens **3** e **4** utilizando esses novos conjuntos.