INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA PIAUÍ	**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ** **Curso: ADS** **Disciplina: Tópicos Especiais em Sistemas de Informação** **Professor: Ely**

Exercícios extras 01

Considerando a base census.csv disponibilizada anteriormente, implemente as seguintes questões:

1) Carregue o dataset em um dataframe e exiba os 5 primeiros registros.
2) Conhecendo o dataset:
    a. Use a função describe e veja os valores mínimos, máximos e médias para os campos age, education-num e hour-per-week.
    b. Crie um dataframe temporário com as 3 colunas acima e exiba histogramas.
    c. Verifique se há valores nulos nas colunas.
    d. Verifique que todos os valores estão com um espaço em branco a após a vírgula no arquivo original. Retire esses espaços no próprio dataframe lido.

        Utilize alguma das respostas encontradas em:

        https://stackoverflow.com/questions/40950310/strip-trim-all-strings-of-a-dataframe.

    e. Exiba os valores únicos para as colunas: education, workclass, education,education-num , marital-status, occupation, relationship, race, sex, native-country e income.

        Dica: atribua a um novo datraframe o resultado do “drop” do dataframe original excluindo as colunas [0, 2, 4, 10,11, 12] e faça um algoritmo interativo para isso para exibir os valores distintos. Há um exemplo nos slides tesi 03

    f. Exiba o total de linhas por valor único obtido no item anterior.
    g. Exiba gráficos de pizza (Pie Charts) para as colunas do item e. usando como referência o código abaixo:

        ```
        import matplotlib.pyplot as plt

        df_temp[coluna].value_counts(sort=False).plot.pie(autopct='%1.1f%%', shadow=True, figsize=(15, 15))

        plt.show()
        ```

Novamente, faça uma iteração para exibir os vários gráficos de pizza.

h. Exiba uma matriz de correlação para o dataframe original. Investigue porque motivo muitas colunas não aprecem na matriz.

3) Separando os dataframes x e y:
   a. Crie o dataframe x utilizando a função iloc:
      x = df.loc[0,'age':'native-country']
   b. Crie o dataframe y:
      y = df['income']
   c. Construa um novo dataframe em x com a conversão para variáveis categóricas (função get_dummies). Verifique quantas novas colunas foram criadas.

4) Utilizando k-folding com k=10, execute os seguintes algoritmos e compare os desempenhos:
   a. MultinomialNB
   b. GaussianNB
   c. AdaBoost
   d. RandomForest

5) Usando a implementação da função cross_val_score não temos acesso aos resultados de forma analítica, ou seja, não temos acesso às previsões de cada linha. Temos apenas os resultados do percentual de acerto. Assim não podemos calcular a matriz de confusão.

Com o uso da função StratifiedKFold, podemos iterar pelo número de “k”s e obter os resultados específicos de cada uma das k interações. Assim, para cada resultado temos uma matriz de confusão.

O código abaixo mostra o exemplo para o nosso estudo de caso:

```
import numpy as np

from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import accuracy_score, confusion_matrix

kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=1)
resultados = []
matrizes = []

x_values = x.values

for indice_treinamento, indice_teste in kfold.split(x,
                                        np.zeros(shape=(x.shape[0], 1))):
    # print('Índice treinamento: ', indice_treinamento, 'Índice teste: ', indice_teste)
    modelo = MultinomialNB()

    modelo.fit(x_values[indice_treinamento], y[indice_treinamento])
    previsoes = modelo.predict(x_values[indice_teste])
```

```
    precisao = accuracy_score(y[indice_teste], previsoes)
    matrizes.append(confusion_matrix(y[indice_teste], previsoes))
    resultados.append(precisao)


matriz_final = np.mean(matrizes, axis=0)
resultados = np.asarray(resultados)
print(resultados.mean())
print(matriz_final)
```

Entenda, execute e analise os resultados da matriz de confusão.

6) Com base na matriz de confusão, calcule as seguinte métricas adicionais, considerando a renda acima de 50k como positivo:
    a. Precisão geral (accuracy).
    b. Precisão.
    c. Recall.

    Use como base para estudo os seguintes links:

- https://medium.com/as-m%C3%A1quinas-que-pensam/m%C3%A9tricas-comuns-em-machine-learning-como-analisar-a-qualidade-de-chat-bots-inteligentes-introdu-57ff30424192
- https://medium.com/as-m%C3%A1quinas-que-pensam/m%C3%A9tricas-comuns-em-machine-learning-como-analisar-a-qualidade-de-chatbots-inteligentes-conceitos-a5b586053973
- https://medium.com/as-m%C3%A1quinas-que-pensam/m%C3%A9tricas-comuns-em-machine-learning-como-analisar-a-qualidade-de-chat-bots-inteligentes-m%C3%A9tricas-1ba580d7cc96
- http://mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/
- https://www.youtube.com/watch?v=85dtiMz9tSo