MBA em Ciência de Dados

Técnicas Avançadas de Captura e Tratamento de Dados

Módulo II - Tratamento de Dados

Avaliação

Moacir Antonelli Ponti

CeMEAI - ICMC/USP São Carlos

As respostas devem ser fornecidas no Moodle. O notebook é apenas para a implementação dos códigos que fornecerão as respostas

---

In [1]:
```
# carregando as bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

from sklearn.svm import SVC
from sklearn.linear_model import Ridge
from sklearn import metrics

# carregando dados
data_orig = pd.read_csv("./dados/pib_mba_avaliacao.csv")
```

Vamos utilizar uma base de dados baixada do IBGE com o PIB per capita para cada município brasileiro, essa base foi modificada para o propósito dos exercícios abaixo. Essa base possui as seguintes colunas:

- gid - identificador geográfico do município
- UF - unidade federativa
- nome - nome do município
- Censo - ano do censo relativo aos dados
- PIB - total do PIB
- Pop_est_2009 - populacao estimada
- PIB_percapita - PIB per capita segundo os dados
- Descrição - Descrição do dados
- classe - classe do município
- desemprego - índice de desemprego na cidade no ano do Censo:

Carregue usando: `pd.read_csv("./dados/pib_mba_avaliacao.csv")`

Antes de iniciar:

1. Inspecione o tipo dos atributos e seus valores possíveis
   A. Verifique se há variáveis irrelevantes para a base de dados, ou que atrapalhem a análise. Identifique-as e remova-as.
   B. realize uma limpeza inicial considerando a:
      a. correção dos dados que forem possíveis inferir o valor verdadeiro, ajustando e padronizando-os. Anote quais variáveis isso ocorreu.
      b. conversão dos atributos que deveriam ser numéricos para numérico - inspecione os valores para garantir que a conversão não vá gerar dados faltantes de forma desnecessária, substituindo por numeros os que forem possíveis como por exemplo o atributo "floor" como visto na aula em que substituímos dados por 0. Anote as variáveis em que isso ocorreu. Verifique ainda se há padronizacao do tipo de dado (separador de decimal por ponto ou vírgula)

         OBS: utilize `df = df.drop('nome_Variavel', 1)` para remover uma variável de um dataframe `df` . Caso queira manter uma cópia por seguranca, utilize `df_copy = df.copy()` para realizar a cópia.

         Importante: nesse passo, ainda não remova outliers!

1. Procure por municípios duplicados, considerando nome e UF. Para isso use:
   `data.duplicated(variaveis)`

1. Remova as linhas duplicadas encontradas no passo anterior, tratando da melhor forma as duplicatas

Questão 1)

Considerando a limpeza inicial realizada, quais variáveis possuiam valores que precisaram ser padronizados ou corrigidos de forma a não causar perda de dados e/ou inconsistências

(a) UF, Desemprego, Censo, gid e nome
(b) UF, Desemprego, Censo
(c) UF e gid
(d) UF e Censo

Questão 2)

Após verificar duplicatas, quantas linhas foram removidas?

(a) 5
(b) 3
(c) 4
(d) 6

Questão 3)

Das 11 colunas iniciais, havia alguma identificada como irrelevante e que foi removida?

(a) não
(b) sim: as colunas 1 e 9
(b) sim: a coluna 9
(d) sim: a coluna 1, 2 e 9

Questão 4)

Vamos analisar possíveis outliers. Utilize o método da análise da dispersão pelo desvio padrão e inspecione as colunas 'gid', 'Censo', 'PIB', 'Pop_est_2009', 'desemprego', procurando por outliers globais com critério de 2 desvios padrões, i.e. 2σ.

Quantos outliers foram encontrados, respectivamente, para 'gid', 'Censo', 'PIB', 'Pop_est_2009' e 'desemprego'?

(a) 0, 0, 1, 5, 27
(b) 0, 5, 9, 44, 1
(c) 0, 5, 5, 27, 1
(d) 0, 5, 5, 44, 2

Questão 5)

Analisando os outliers retornados em 'Censo' e 'desemprego' na questão anterior, quantos valores respectivamente, parecem ser outliers globais verdadeiros e para os quais se recomenda remover o valor antes de qualquer análise posterior?

(a) 5, 0
(b) 0, 1
(c) 5, 1
(d) 1,1

Questão 6)

Utilize a base de dados após a limpeza inicial (sem remover outliers). Imprima o total de valores faltantes em cada variável.

Quais variáveis possuem valores faltantes e em qual número?

(a) nome: 2, Censo: 6, PIB: 1, PIB_percapita: 4, desemprego: 30
(b) nome: 1, Censo: 6, PIB_percapita: 4, desemprego: 30
(c) nome: 2, Censo: 7, PIB: 1, PIB_percapita: 4, desemprego: 30
(d) nome: 2, Censo: 6, PIB_percapita: 4, desemprego: 30

Questão 7)

Codifique uma funcão que preencha valores faltantes de variáveis numéricas utilizando a média condicionada (ou agrupada) a uma outra variável C da base. Essa funcao deverá:

1. calcular a média da variável alvo A (a ser preenchida) relativa a (ou agrupada por) cada valor distinto da variável categórica selecionada C
2. atribuir a média calculada de forma agrupada a todas as linhas cuja variável alvo é faltante e que possua o valor da variável categórica correspondente
3. o valor atribuido deve seguir o mesmo tipo da variável alvo, ou seja, int, float, etc. Quando int, realize o arredondamento utlizando `np.round(,0)` , quando float64 utilize `np.round(,1)`

Use a funcão para preencher dados faltantes de desemprego condicionado a UF. Considerando arredondamento para 4 casas decimais, qual é a média da coluna desemprego para todas as linhas, antes e depois de realizar o preenchimento?

(a) antes: 6.6406, depois: 6.6423
(b) antes: 6.6406, depois: 6.6406
(c) antes: 6.6423, depois: 6.6406
(d) antes: 6.6423, depois: 6.6423

Questão 8)

Considere o atributo 'classe' apenas para a UF 'Rio de Janeiro' e analise a distribuicao dos seus valores.

Para executar um algoritmo de aprendizado em que a o atributo alvo é 'classe', qual seria a abordagem mais indicada:

(a) considerar cenário desbalanceado com 2 classes minoritárias e estudar medidas para compensar esse desbalanceamento
(b) realizar análise com os dados originais, mesmo desbalanceados
(c) considerar cenário desbalanceado com 1 classes minoritárias e estudar medidas para compensar esse desbalanceamento
(d) considerar cenário desbalanceado com 4 classes minoritárias e estudar medidas para compensar esse desbalanceamento