MBA em Ciência de Dados

Técnicas Avançadas de Captura e Tratamento de Dados

Módulo III - Aquisição e Transformação 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

---

Utilize as bibliotecas e carrege os dados conforme descrito abaixo

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

Questão 1)

Porque é importante auditar fontes de dados já existentes e utilizar método e planejamento para realizar coleta de dados?

(a) para evitar analisar dados com viés e chegar a conclusões inválidas
(b) para obter acurácias maiores
(c) para evitar analisar dados errôneos e assim obter estimadores com menos erro
(d) para evitar usar evidências anedotais para chegar a conclusões

Questão 2)

Acesse o portal : http://catalogo.governoaberto.sp.gov.br/ (http://catalogo.governoaberto.sp.gov.br/)

Procure pelo arquivo CSV relativo a "Quantidade de alunos por tipo de ensino da rede estadual - 01/2019" (Secretaria da Educação - Sede)

Carregue os dados (considere as particularidades do arquivo e não carregue o cabeçalho). Depois, remova as colunas 21 em diante, mantendo as colunas de 0 a 20. Essas colunas restantes possuem significado de acordo com o "dicionário de dados" disponível ao visualizar o recurso dos dados. Sendo rotuladas da seguinte forma:

- CDREDE
- DE
- CODMUN
- MUN
- CATEG
- COD_ESC
- TIPOESC
- CODVINC
- NOMESC
- ENDESC
- NUMESC
- BAIESC
- EMAIL
- FONE 1
- ZONA
- ED_INFANTIL
- CLASSES ESPECIAIS
- SALA DE RECURSO
- ANOS INICIAIS
- ANOS FINAIS
- ENSINO MEDIO

Após carregar os dados e nomear as colunas, crie um novo atributo 'TOT' por meio da soma das 6 últimas colunas: ED_INFANTIL, CLASSES ESPECIAIS, SALA DE RECURSO, ANOS INICIAIS, ANOS FINAIS e ENSINO MEDIO.

Atribua nulo (nan) aos elementos cujo total (TOT) é zero.

Realize a **discretização** da variável 'TOT' utilizando:

1. o método dos quantis, utilizando 5 valores alvo, relativos aos quantis 0, 20, 40, 60, 80, 100
2. o método dos intervalos, utilizando 5 intervalos alvo: (0, 10] (10, 50] (50, 100] (100, 500] (500, max(TOT)], em que max(TOT) é o maior valor desse atributo

Use o método qcut() para o item 1 e cut() para o item 2

Adicione essa novas variáveis na base, com os nomes 'TOT_5Q' (quantis) e 'TOT_5I' (intervalos)

Qual é a quantidade de dados (frequências dos valores discretizados) na base após a discretização, relativos ao primeiro intervalo e ao último intervalo para, respectivamente, TOT_5Q e TOT_5I?

(a) TOT_5Q: primeiro 0.999, último 2269. TOT_5I: primeiro 0, último 2269
(b) TOT_5Q: primeiro 1051, último 1038. TOT_5I: primeiro 3099, último 18
(c) TOT_5Q: primeiro 0.999, último 323. TOT_5I: primeiro 500, último 2269
(d) TOT_5Q: primeiro 1051, último 1041. TOT_5I: primeiro 18, último 3099.

Questão 3)

Normalize 2 variáveis da base: TOT e ENSINO MEDIO

- ENSINO MEDIO utilizando normalização da média
- TOT utilizando normalização L-2

Para isso, codifique funções que recebam uma coluna por parâmetro e retornem um atributo já normalizado

Depois, aplique as funções e crie novas variáveis com os atributos normalizados: MEDIO_nm e TOT_l2.

Após normalização, quais os valores de média, desvio padrão, mínimo e máximo dessas variáveis, arredondando para 3 casas decimais?

(a) MEDIO_nm: 0.000, 0.132, 0.000, 0.876; TOT_l2: 0.012, 0.007, 0.000, 1.000
(b) MEDIO_nm: 0.000, 0.132, -0.124, 0.876; TOT_l2: 0.012, 0.007, 0.000, 0.043
(c) MEDIO_nm: 1.000, 0.132, 0.000, 2.000; TOT_l2: 0.012, 0.007, 0.000, 1.000
(d) MEDIO_nm: 0.000, 0.500, -1.124, 1.125; TOT_l2: 0.012, 0.007, 0.000, 0.043

Questão 4)

Utilize os atributos 'ENSINO MEDIO' e 'ANOS FINAIS'. Vamos transformá-los por meio da função logaritmica. Para isso:

1. Faça uma cópia da base de dados, e atribua nulo (nan) a todos os valores iguais a zero nesses atributos,
2. Transforme esses atributos utilizando a operação da raiz quadrada e os adicione à base de dados com novos nomes, ex. sqrt(ENSINO MÉDIO) e sqrt(ANOS FINAIS),
3. Remova todas as linhas que possuam nulo (nan) em qualquer um dos atributos transformados,
4. Utilizando x como sqrt(ENSINO MÉDIO) e y como sqrt(ANOS FINAIS), ajuste um modelo de regressão linear entre as duas variáveis. Use `LinearRegression()` do `sklearn` . Aqui não estamos interessados em separar treinamento e teste, só queremos ajustar uma função linear das duas variáveis, considerando todo os dados disponíveis.
5. Obtenha (imprima) o coeficiente da regressão linear aprendido, disponível no atributo coef_ do modelo inferido no passo anterior.

Qual o valor obtido do coeficiente no passo 5 acima, arredondado para 3 casas decimais?

(a) 0.481
(b) 0.668
(c) 0.690
(d) 0.657

Questão 5)

Utilize a base de dados antes da modificação feita na Questão 4. Codifique as variáveis categóricas 'MUN' (categórica nominal) e 'TOT_5I' (categórica ordinal).

Para MUN use números inteiros sequenciais, iniciados por 0 para codificar a variável em ordem alfabética, e gere um novo atributo MUN_cod

Para TOT_5I, use números inteiros sequenciais, iniciado por 0 para codificar a variável segundo sua ordenação, utilize para isso a ordenação crescente de seus valores únicos. Gere um novo atributo TOT_5I_ord.

A seguir, use a função `value_counts()` para mostrar a frequência de cada código na base de dados. Responda abaixo quais valores dos novos atributos (após codificação realizada) possuem a maior frequência (maior contagem):

(a) MUN_cod: código 1124; TOT_5I: código 3095
(b) MUN_cod: código 563; TOT_5I: código 3
(c) MUN_cod: código 1124; TOT_5I: código 4
(d) MUN_cod: código 563; TOT_5I: código 4