República Federativa do Brasil

Ministério da Educação

Universidade Federal do Amazonas

Instituto de Computação

Ciência de Dados

Trabalho Prático I: Exploração de Dados

Professor Dr. Marco Cristo

Alunos: `Nome Sobrenome, Nome Sobrenome e Nome Sobrenome`
E-mail: `nome.sobrenome@icomp.ufam.edu.br, nome.sobrenome@icomp.ufam.edu.br e nome.sobrenome@icomp.ufam.edu.br`

Preliminares

In [1]:
```
# Imports
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from scipy import stats

import sklearn
import sys
import os
```

In [2]:
```
print("Python", sys.version)
print("--------")
print("Pandas:", pd.__version__)
print("Numpy:", np.__version__)
print("SKLearn:", sklearn.__version__)
```

```
Python 3.11.6 (v3.11.6:8b6ee5ba3b, Oct  2 2023, 11:18:21) [Clang 13.0.0 (clang-1300.0.29.30)]
--------
Pandas: 2.2.0
Numpy: 1.26.4
SKLearn: 1.7.1
```

In [3]:
```
DATA_PQT = 'data/enem2019.parquet'
MUNICIPIO_PROVA = 'Londrina' # Use o município que foi especificado para sua equipe.
```

In [4]:
```
# Caminho do arquivo
df_enem = pd.read_parquet(DATA_PQT)
df_enem.shape
```

Out[4]: (599982, 56)

Qualidade de dados

1 - Quais colunas possuem campos nulos em `df_enem` ?
Dica: Há muitas formas de fazer isso em Pandas, como por exemplo através da utilização do método `isnull()`.

In [ ]:
```

```

2 - Remova todas as instâncias que não possuem valores para as colunas `NO_MUNICIPIO_NASCIMENTO` ou `SG_UF_NASCIMENTO` Para as colunas `CO_PROVA_CN` , `CO_PROVA_CH` , `CO_PROVA_LC` , `CO_PROVA_MT` e `TP_STATUS_REDACAO` , remova apenas as instâncias onde todos esses valores encontram-se ausentes.
Dica: Os argumentos `subset` e `how`, do método `dropna()` do Pandas podem lhe ajudar.

In [ ]:
```

```

3 - Substitua dados faltantes por valores razoáveis. Valores ausentes nas colunas `TP_ENSINO` e `TP_STATUS_REDACAO` podem ser substituídos por 0 (0 = Não informado). No caso das notas, se elas forem NaN, podemos substituí-las por 0 (estamos penalizando quem faltou uma prova, mas fez as demais, em parte porque há poucos casos assim nesta base). Já os códigos do tipo de prova (`CO_PROVA_**`), quando ausentes, podem ser substuídos para 500 (_500 = Faltou à prova_). Após isso, imprima o vetor resultante para confirmar se não há mais valores NaNs.

In [ ]:

Análise e visualização de dados

4 - Considerando a distribuição de notas em sua cidade de estudo, quem se saiu melhor: alunos das escolas de ensino médio públicas ou privadas? O mesmo se observa para o resto do país? Finalmente, para que provas, o desempenho das escolas públicas da sua cidade de interesse é *significativamente* pior (confiança de 95%) que a observada nas escolas do resto do país? *Dica*: boxplots são ótimos para comparar distribuições. Você pode verificar relevância da diferença nas notas usando um teste não paramétrico (ex: mannwhitneyu) unicaudal.

In [ ]:

4.1 - Explique os resultados obtidos.

5 - Como se comparam os inscritos que fizeram a prova com intuito de apenas treinar seus conhecimentos com os demais? E o que se observa para o resto do país? Finalmente, para que provas de sua cidade de interesse, o desempenho dos alunos treinando é significativamente melhor (confiança de 95%) que o dos alunos fazendo a prova para valer? Novamente considere as distribuições de notas, mas desta vez use gráficos de violino e avalie diferenças de notas com o teste T de Student.

In [ ]:

5.1 - Explique os resultados obtidos.

6 - Sobre a redação, como se comparam os desempenhos dos inscritos conforme o ano de conclusão do ensino médio em sua cidade de interesse? E no resto do país? Usando um gráfico de linhas, trace as notas médias obtidas contra o ano de conclusão. *ATENÇÃO*: Em `TP_ANO_CONCLUIU`, defina 2019 para todas as intâncias onde `TP_ST_CONCLUSAO` seja igual a 2 (2 = "Estou cursando e concluirei o Ensino Médio em 2019").

In [ ]:

6.1 - Explique os resultados obtidos.

7 - Qual a relação entre o grau de instrução dos pais (`Q001` e `Q002`) e o desempenho dos seus filhos, tanto na sua cidade de interesse quanto no resto do país? Considere a média de todas as notas do candidato como seu desempenho.

In [ ]:

7.1 - Explique os resultados obtidos.

Ainda considerando o desempenho calculado na questão anterior e, agora, as seguintes faixas de renda (`Q006`):

A = Nenhuma renda.
B = Até R\$ 998,00.
C = De R\$ 998,01 até R\$ 1.497,00.
D = De R\$ 1.497,01 até R\$ 1.996,00.
E = De R\$ 1.996,01 até R\$ 2.495,00.
F = De R\$ 2.495,01 até R\$ 2.994,00.
G = De R\$ 2.994,01 até R\$ 3.992,00.
H = De R\$ 3.992,01 até R\$ 4.990,00.
I = De R\$ 4.990,01 até R\$ 5.988,00.
J = De R\$ 5.988,01 até R\$ 6.986,00.
K = De R\$ 6.986,01 até R\$ 7.984,00.
L = De R\$ 7.984,01 até R\$ 8.982,00.

M = De R\$ 8.982,01 até R\$ 9.980,00.
N = De R\$ 9.980,01 até R\$ 11.976,00.
O = De R\$ 11.976,01 até R\$ 14.970,00.
P = De R\$ 14.970,01 até R\$ 19.960,00.
Q = Mais de R\$ 19.960,00.

Responda:

8 - É possível afirmar que, dos inscritos que concluem o ensino médio em 2019, os mais ricos são os que obtém o maior desempenho, tanto na sua cidade de interesse quanto no resto do país? Neste caso, além de uma inspeção visual, prove sua afirmação usando um teste de correlação de Spearman.

In [ ]:

8.1 - Explique os resultados obtidos.

9 - Qual a nota geral média dos alunos, de acordo com o tipo de estado civil (`TP_ESTADO_CIVIL`), tanto em sua cidade de interesse quanto no resto do país?

In [ ]:

9.1 - Explique os resultados obtidos.

10 - Qual a nota geral média dos alunos, de acordo com o tipo da lingua estrangeira escolhida (`TP_LINGUA`) tanto em sua cidade de interesse quanto no resto do país?

In [ ]:

10.1 - Explique os resultados obtidos.

11 - Qual a nota geral média dos alunos, de acordo com ocupação dos pais/responsáveis e (`Q003` e `Q004`) tanto em sua cidade de interesse quanto no resto do país?

In [ ]:

11.1 - Explique os resultados obtidos.

Transformação e Engenharia de atributos

12 - A partir do campo `NU_IDADE` , que armazena a idade dos inscritos, crie a coluna `FAIXA_ETARIA` com os seguintes intervalos:

- [10..18)
- [18..22)
- [22..26)
- [26..33)
- [33..40)
- [40..55)
- [55..65)
- [65..70)
- [70..100]

ATENÇÃO: Apague instâncias com idade inferior a 10 ou superior a 100 anos.

In [ ]:

Considerando apenas os alunos que não faltaram a nenhuma prova (`TP_PRESENCA_**`), queremos tentar prever se um dado inscrito irá obter uma nota de redação acima ou abaixo da média, a partir das seguintes informações:

- `FAIXA_ETARIA` , `NO_MUNICIPIO_RESIDENCIA` , `TP_SEXO` , `TP_ESTADO_CIVIL` , `TP_ST_CONCLUSAO` , `TP_ANO_CONCLUIU` , `TP_ESCOLA`
- `CO_PROVA_CH` , `CO_PROVA_LC` , `NU_NOTA_CH` , `NU_NOTA_LC` , `TP_LINGUA`
- `Q001` , `Q002` , `Q003` , `Q004` , `Q005` , `Q006` , `Q021` , `Q022` , `Q024` , `Q025`
- `TARGET`

13 - Prepare os dados para aprendizado, convertendo strings para dados categóricos, mantendo sem mudanças os dados categóricos não numéricos e padronizando dados numéricos usando Z-score.
ATENÇÃO: Você deve criar a coluna `TARGET` para ser utilizada como alvo de classificação e avaliação de atributos. Ela deve ser `1`, caso a nota da redação do aluno seja maior que a média, ou `0`, caso contrário.

In [ ]:

14 - Que atributos são os mais importantes considerando correlação absoluta com o alvo, ANOVA, chi quadrado, Informação Mútua e a importância das variáveis de acordo com uma floresta aleatória? Considere apenas candidatos em sua cidade de interesse.
Dica: Para facilitar a visualização, coloque todas as métricas em um único dataframe. Então as normalize para que todas fiquem entre 0 e 1. Obtenha então uma média agregada das métricas normalizadas e use esta média agregada para ordenar as variáveis da mais à menos importante.

In [ ]: