Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

18 de janeiro de 2023

Lista 3: Manipulação e modelagem de dados com Spark

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 2

Prof. Guilherme Rodrigues

César Augusto Fernandes Galvão (aluno colaborador)

Gabriel Jose dos Reis Carvalho (aluno colaborador)

1. **As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Rmarkdown* ou outra ferramenta equivalente.**
2. **O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, *blogs* e artigos.**
3. **O trabalho é individual. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.**
4. **Os códigos *R* utilizados devem ser disponibilizados na integra, seja no corpo do texto ou como anexo.**
5. **O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.**
6. **O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.**
7. **Escreva seu código com esmero, evitando operações redundantes, visando eficiência computacional, otimizando o uso de memória, comentando os resultados e usando as melhores práticas em programação.**

Questão 1: Criando o cluster spark.

a) Crie uma pasta (chamada datasus) em seu computador e faça o download dos arquivos referentes ao Sistema de informação de Nascidos Vivos (SINASC), os quais estão disponíveis em https://datasus.saude.gov.br/transferencia-de-arquivos/.

Atenção: Considere apenas os Nascidos Vivos no Brasil (sigla DN) entre 1994 e 2020, incluindo os dados estaduais e excluindo os arquivos referentes ao Brasil (sigla BR). Use wi-fi para fazer os downloads!

Dica: O endereço ftp://ftp.datasus.gov.br/dissemin/publicos/SINASC/1996_/Dados/DNRES/ permite a imediata identificação dos endereços e arquivos a serem baixados.

b) Usando a função `p_load` (do pacote `pacman`), carregue os pacotes `arrow` e `read.dbc` e converta os arquivos baixados no item a) para formato o *.parquet*. Em seguida, converta para *.csv* apenas os arquivos referentes aos estados GO, MS e ES. Considerando apenas os referidos estados, compare o tamanho ocupado pelos arquivos nos formatos *.parquet* e *.csv* (use a função `file.size`).

c) Crie uma conexão `Spark`, carregue para ele os dados em formato *.parquet* e *.csv* e compare os respectivos tempos computacionais. Se desejar, importe apenas as colunas necessárias para realizar a Questão 2.

OBS: Lembre-se de que quando indicamos uma pasta na conexão, as colunas escolhidas para a análise precisam existir em todos os arquivos.

Questão 2: Preparando e modelando os dados.

Atenção: **Elabore seus comandos dando preferência as funcionalidades do pacote sparklyr**.

a) Faça uma breve análise exploratória dos dados (tabelas e gráficos) com base somente nas colunas existente nos arquivos de 1996. O dicionário das variaveis encontra-se no mesmo site do item a), na parte de documentação. Corrija eventuais erros encontrados; por exemplo, na variavel sexo são apresentados rótulos distintos para um mesmo significado.

b) Ultilizando as funções do **sparklyr**, preencha os dados faltantes na idade da mãe com base na mediana. Se necessário, faça imputação de dados também nas demais váriaveis.

c) Novamente, ultilizando as funções do **sparklyr**, normalize (retire a média e divida pelo desvio padrão) as variáveis quantitativas do banco.

d) Crie variáveis dummy (*one-hot-encoding*) que conjuntamente indiquem o dia da semana do nascimento (SEG, TER, ...). Em seguida, *binarize* o número de consultas pré-natais de modo que "0" represente "até 5 consultas" e "1" indique "6 ou mais consultas". (Ultilize as funções **ft_**)

e) Particione os dados aleatoriamente em bases de treinamento e teste. Ajuste, sobre a base de treinamento, um modelo de regressão logistica em que a variável resposta (y), indica se o parto foi ou não cesáreo. Analise o desempenho preditivo do modelo com base na matrix de confusão obtida no conjunto de teste.