Desafio para o processo seletivo Murabei

Murabei Data Science

1. Objetivo do Desafio

O objetivo do desafio é avaliar a organização do código e a capacidade de executar tarefas simples de tratamento e modelagem de dados. Os dados do desafio consistem em características de escolas e alunos, sendo que o objetivo da modelagem é prever o resultado no exame normalizado (normexam) de acordo com as características dos alunos e das escolas.

O modelo deve ser capaz de responder quais do tipo de escola (gênero misto, só de homens ou só de mulheres) é mais eficiente na formação de seus alunos de acordo com o resultado do exame normalizado. Caso não seja possível terminar o desafio, pode ser entregue uma análise parcial dos dados com os passos que foram problemáticos.

Pode ser utilizado R ou Python como linguagens de programação.

2 Base de dados

A base de dados fornecida foi extraída do curso de modelagem de Harvard[1] e está organizada em quatro arquivos diferentes:

cat school data.csv Variáveis categóricas que são aplicadas às escolas.

num school data.csv Variáveis numéricas que são aplicadas às escolas.

cat student data.csv Variáveis categóricas que são aplicadas aos estudantes.

num student data.csv Variáveis numéricas que são aplicadas aos estudantes.

Os dados das diferentes bases são descritos abaixo. As tabelas de estudante e escola são cruzadas através do campo school, os ids dos estudantes (student) são únicos para cada escola.

variavel,	descrição
school	School ID.
normexam	Normalized exam score.
schgend	School gender.
schavg	School average of intake score.
vr	Student level Verbal Reasoning (VR) score band at intake.
intake	Band of student's intake score.
standLRT	Standardised LR test score
sex	Sex of the student - levels are 'F' and 'M'.
type	School type.
student	Student id (within school).

3 Resultados esperados

A resposta do desafio deve ser entregue por e-mail em resposta ao e-mail do próprio desafio, com o arquivo de tratamento e modelagem em anexo. No final do arquivo em comentário, deve ser colocada a resposta sobre a eficiência dos diferentes tipos de escola, justificada pelos parâmetros do modelo (coisa curta, de um parágrafo apenas).

O nome do arquivo em anexo deve conter o nome do candidato, assim com um cabeçalho, para facilitar a nossa organização dos códigos.

Obrigado por participar do processo e boa sorte!

[1] http://tutorials.iq.harvard.edu/R/Rstatistics/Rstatistics.html