Propósito

Este desafio tem como objetivo avaliar sua capacidade de resolver problemas atuando como um cientista de dados. Portanto, as técnicas, algoritmos, teorias, estrutura e qualidade do código vão dizer muito sobre você.
Este desafio avaliará as seguintes habilidades:

- Entendimento do problema e capacidade de abstração.
- Modelagem de dados de diferentes naturezas utilizando algoritmos de agrupamento.
- Execução e avaliação do modelo criado.
- Reprodutibilidade do modelo criado.
- Estrutura e qualidade do código.
- Forma da apresentação dos resultados.

Sobre o Desafio

A Easynvest é uma corretora de valores que ajuda milhares de clientes a investir seu dinheiro de uma forma fácil e ágil. Nossos produtos oferecem uma plataforma online e mobile de negociação de títulos públicos entre pessoas físicas e o programa do Tesouro Direto do governo.

Todo dia milhares de pessoas se tornam nosso clientes e a variabilidade de suas características é impressionante. São pessoas com características parecidas ou pessoas totalmente diferente.

Seu trabalho consiste em:
1) Agrupar os usuários, encontrando grupos bem definidos com características comuns.
2) Justificar o algoritmo de clusterização utilizado.
3) Apresentar métricas de performance do algoritmo utilizado.
4) Expor métricas de performance para avaliar os clusters obtidos.
5) Explicar os resultados.

Stack de Tecnologias

- Linguagem de programação: R ou Python.
- Visualização: Pacotes do R/Python ou ferramentas de visualização open source.

Não há restrições para uso de bibliotecas que possam otimizar o seu código. ***But, be careful!*** Você pode utilizar pacotes com os algoritmos já implementados ou você pode usar algoritmos implementados por você.

Dataset

O dataset para essa desafio possui as seguintes características:

1. **ID (discreta)**: Identificador único
2. **GEO_REFERENCIA (discreta)**: Identificador geográfico
3. **DATA_NASCIMENTO (discreta)**: data de nascimento
4. **PROFISSAO (categórica)**: profissão do usuário
5. **GENERO (categórica)**: gênero do usuário
6. **ESTADO_CIVIL (categórica)**: estado civil do usuário
7. **VALOR_01 (continua)**: valor 01
8. **VALOR_02 (contínua)**: valor 02
9. **VALOR_03 (contínua)**: valor 03
10. **VALOR_04 (contínua)**: valor 04
11. **PERFIL (categórica)**: perfil do cliente

O dataset está disponível nesse link aqui.

Não basta apenas você saber executar o projeto

É importante que seu projeto tenha uma documentação consistente que permita a qualquer pessoa executar ou modificar o modelo criado.

Dica: ter o seu projeto em um repositório online é uma boa ideia, afinal de contas sempre trabalhamos em time.

O que esperamos ver ao final?

Nosso time está bastante curioso para ver e analisar o seu modelo. E queremos compartilhar com você alguns dos pontos que avaliaremos no seu projeto:

1. **Interpretação e metodologia:** mostrar de forma clara, por meio da estrutura do código e da documentação, a estratégia adotada para a resolução do problema, bem como todas as premissas assumidas e suas razões.
2. **Modelagem:** explicação da escolha do algoritmo, pontos fracos e fortes perante aos dados e também a outras técnicas.
3. **Performance:** Apresentar métricas de avaliação do modelo. Os clusters são bem definidos? Os clientes são parecidos entre si e diferente dos outros?
4. **Desempenho**: escreveu um código que tem uma boa performance? Não esperamos a solução ótima, mas é interessante saber identificar pontos de melhoria e otimização.
5. **Manutenibilidade**: é um código legível e de fácil manutenção. Segue premissas claras de padronização de código.
6. **Visualização:** Apresentação dos resultados de forma clara e de fácil entendimento através de gráficos, tabelas, web applications, relatórios, entre outros.

O tempo estimado de criação desse projeto é de 48 horas.

BOA SORTE! =)