# Product Data Science - Data Analytics

## Teste de experiência na posição

### Descrição do problema

Você é um funcionário da OMS que deve avaliar os níveis de contaminação de um vírus em um determinado país. As pessoas dentro de uma sociedade podem estar conectadas de alguma maneira (familia, amizade ou trabalho) e cada pessoa possui um conjunto de atributos.

Este vírus afeta esta sociedade como descrito a seguir:

- a taxa de contaminação varia de pessoa para pessoa;
- a taxa de contaminação de uma pessoa A para B é diferente de B para A e depende das características de ambas as pessoas (A e B);
- a contaminação só passa através de indivíduos conectados;
- não existe cura para essa doença;

### O desafio

Foram coletados os dados de contaminação (ou seja, as taxas de contaminação) para metade desta sociedade. Neste problema, você deverá estimar a taxa para o restante dessa sociedade e decidir políticas de saúde com base nos resultados obtidos.

Observação: Para determinar as taxas de contaminação, devem ser levados em consideração tanto as características dos infectados quanto dos infectantes.

### Entregáveis

1. Uma apresentação com os principais resultados da sua análise empírica e as recomendações de políticas de saúde que devem ser baseadas nos dados e nos resultados de sua análise.
2. O código usado, contendo uma breve descrição de como utilizar/rodar e uma justificativa da metodologia utilizada (podem ser feitos gráficos, tabelas ou qualquer análise que ajude a explicar melhor a solução).

### Detalhes da base de dados

Para o desenvolvimento e resolução do problema considere os dois arquivos CSV:

- individuos_espec.csv - contém características de cada indivíduo
    - name: Id dos indivíduos
    - idade: idade dos indivíduos
    - estado_civil: Estado civil dos indivíduos

    - qt_filhos: quantidade de filhos dos indivíduos
    - estuda: caso estudem
    - trabalha: caso trabalhem
    - pratica_esportes: caso pratiquem esportes
    - transporte_mais_utilizado: qual o transporte mais utilizado
    - IMC: valor do índice de massa corporal dos indivíduos
- conexoes_espec.csv - lista das conexões e algumas características das mesmas
    - V1: id do individuo (relação entre os indivíduos)
    - V2: id do individuo (relação entre os indivíduos)
    - grau: familia, amigos, trabalho
    - proximidade: mora_junto, visita_frequente, visita_casual, visita_rara
    - prob_V1_V2: taxa de contaminação de V1 (doente) para V2 (saudável)

## Critérios de avaliação

Iremos avaliar o projeto da seguinte maneira, do maior para o menor em termos de relevância:

1. Construção do pipeline de modelagem.
2. Qualidade da documentação
3. Clareza e capacidade analítica nos relatórios.
4. Qualidade e arquitetura do código, por exemplo:
    - Modularização
    - Funções
    - Testes
    - Logging
5. Reprodutibilidade e instruções para o uso, por exemplo:
    - Docker
    - Conda
    - CLI
    - Virtualenv
6. Escalabilidade de código (processamento) não é um obrigatório, mas será um plus.

O projeto pode ser desenvolvido em inglês ou português e em um repositório GitHub/GitLab. O repositório pode ser aberto ou privado (se for privado, terá que adicionar um de nossos integrantes como Master). Não iremos avaliar você até o final do período de 5 dias (contando a partir da data combinada entre você e a equipe de recrutamento), mas gostaríamos de acompanhar o desenvolvimento do trabalho.

Github para adicionar:
https://github.com/Igorcortez
https://github.com/jhosoume
https://github.com/vhdeluca

## Por último, mas não menos importante:

- Não existe apenas uma resposta certa: estamos procurando um profissional criativo capaz de fornecer uma solução eficiente para este problema.
- A linguagem de programação mais usada em nossa equipe é Python, mas fique à vontade para usar outra de sua preferência.