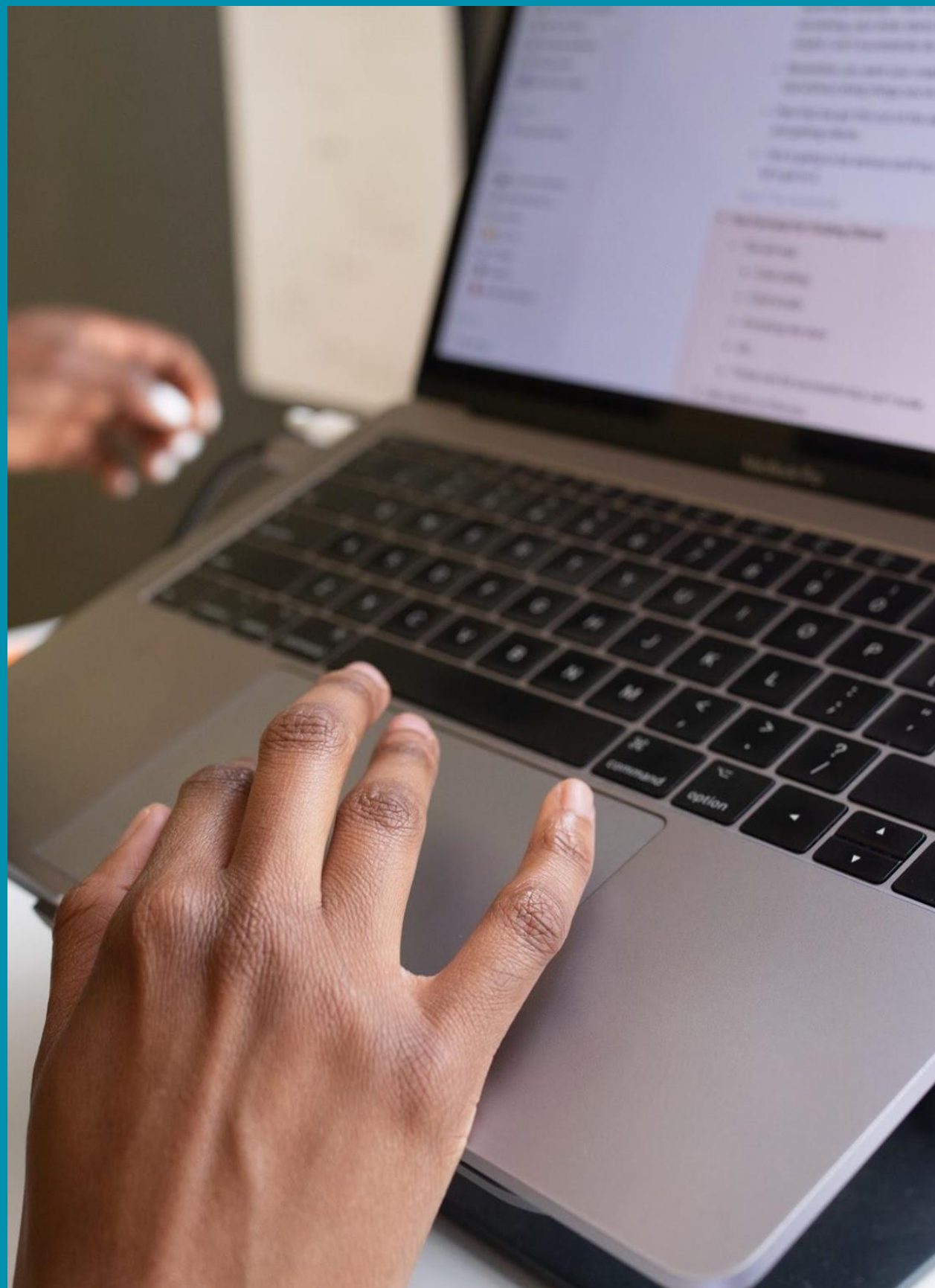


# Ciência de Dados

Teste para o processo seletivo





# O TESTE

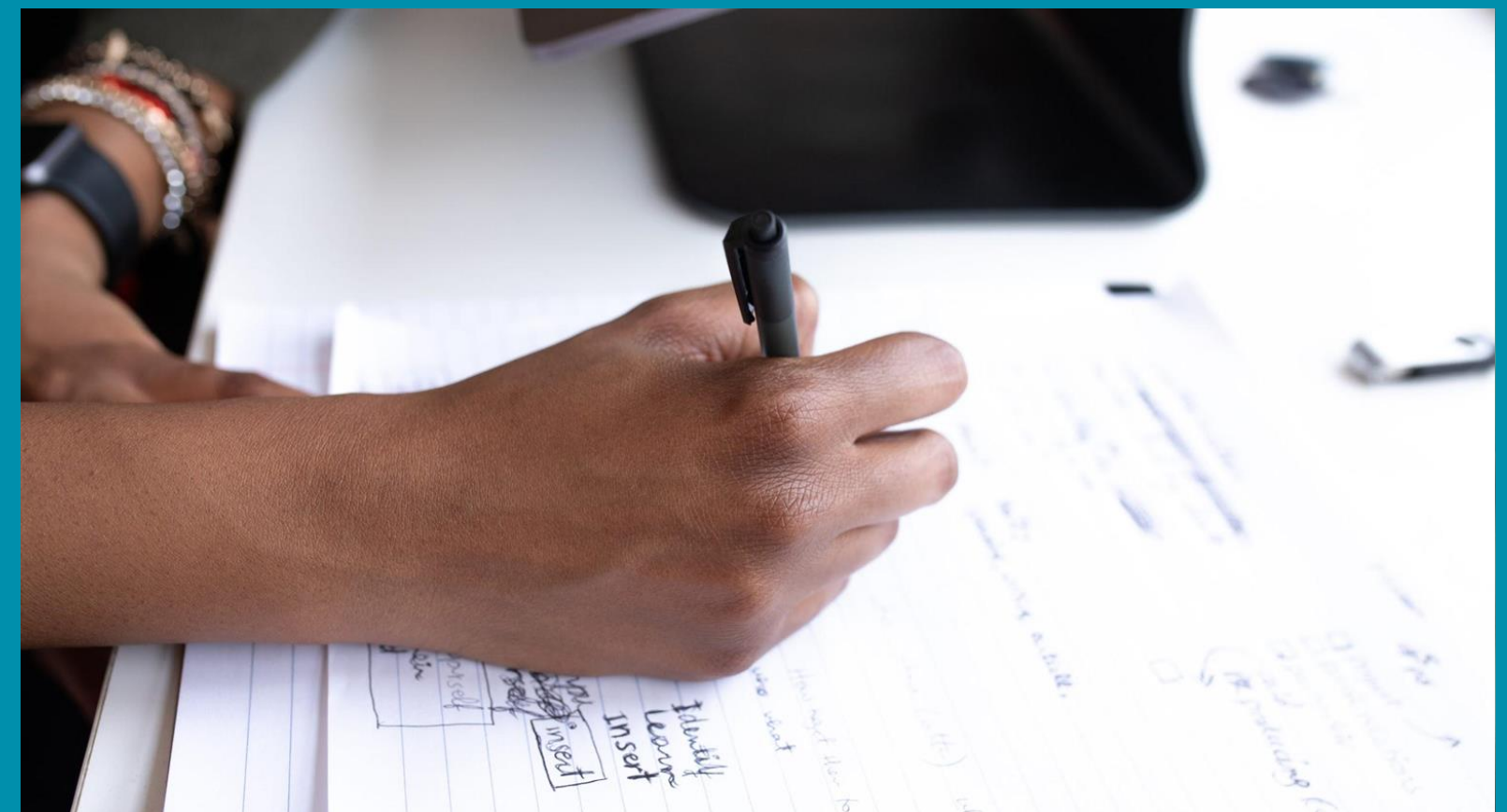
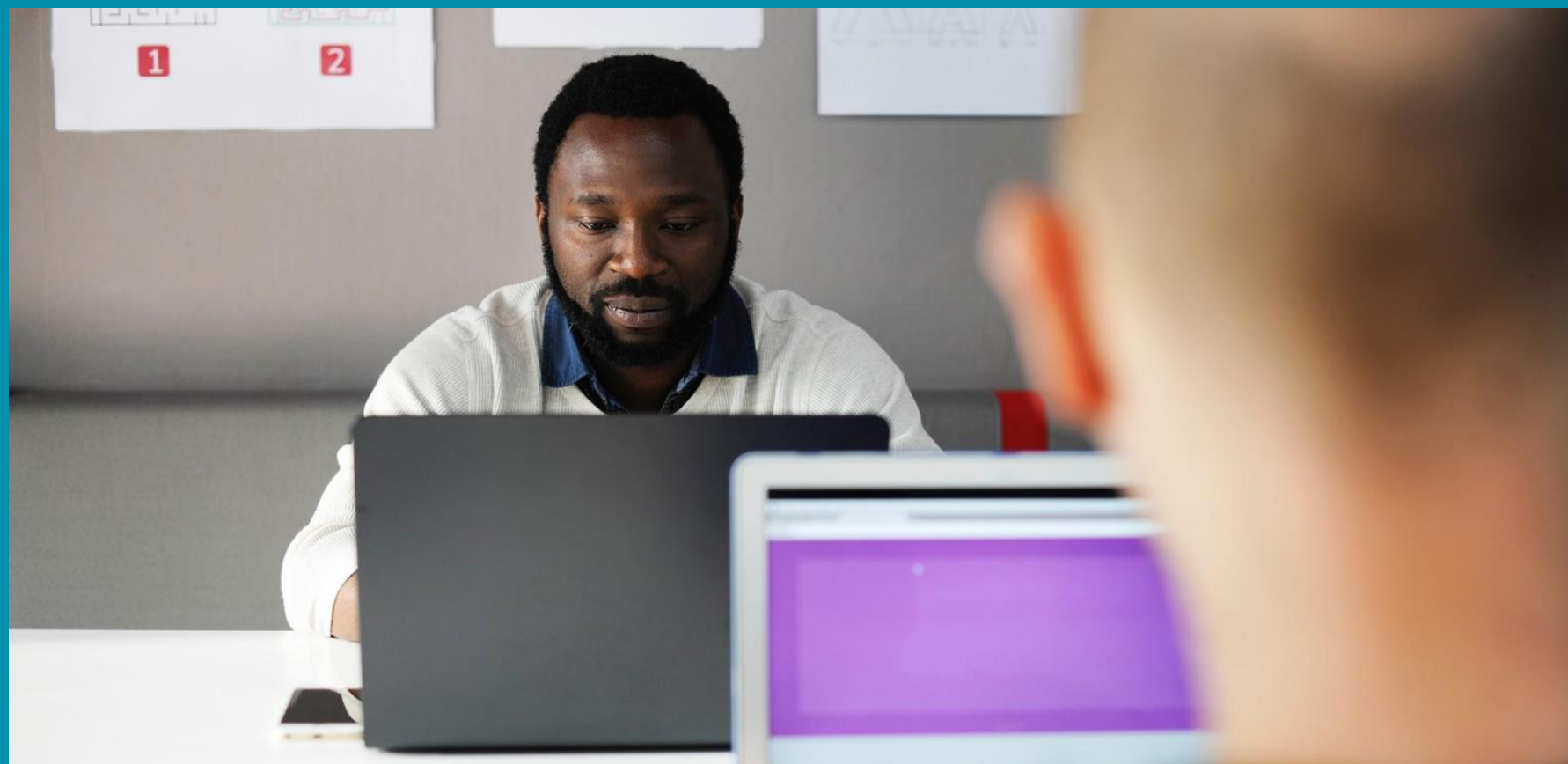
Olá, Candidato!

Esse teste visa testar suas habilidades e conhecimentos sobre ciência de dados. Ele é composto de 2 etapas: questionário técnico com 2 perguntas (a serem respondidas nesse mesmo arquivo) e um desafio.

Este teste pode ser respondido nas seguintes linguagens: R, Python, Scala e SQL

# ETAPA 1

## QUESTIONÁRIO TÉCNICO



# 1 – Visualização de dados



Em 2017 a indústria de games faturou mais que a indústria de música e a indústria de filmes (mídia física) juntos! É um fato que esse nicho do entretenimento conquista cada vez mais pessoas ano após ano. No Brasil mais de 66% disseram que consomem entretenimento do nicho de games.

Anexa você irá encontrar uma base de dados sobre o ranking de vídeo games nos últimos anos e com esse dataset você deverá apresentar as seguintes visualizações:

- a) Histogramas de quantos jogos cada gênero possui nos primeiros 150 títulos do rank
- b) Um gráfico de dispersão entre o ano da publicação e o total de vendas da Nintendo nos últimos 10 anos
- c) As 5 maiores “publishers” em vendas nos Estados Unidos

Obs: Os valores de venda estão em milhões de dólares

# 2 – Machine Learning



- a) O qual a diferença entre um aprendizado de máquina supervisionado e um não supervisionado?  
Dê exemplos de algoritmos para cada um.
- b) Após a execução de dois algoritmos diferentes de machine learning foram resultadas as seguintes matrizes de confusão:

Matriz Modelo 1  
[5740, 519]  
[1119, 9413]

Matriz Modelo 2  
[6751, 705]  
[2005, 7330]

Analisando apenas essas matrizes qual modelo você considera o melhor para ser utilizado? Justifique sua resposta.





# DESAFIO

Vamos supor que a Eleflow foi contratada pela empresa Netflix e ela deseja um modelo preditivo para prever as notas de filmes, para assim decidir se vale a pena ou não colocar esse filme no catálogo. Você daria conta desse desafio?



# INSTRUÇÕES

- Utilize a base de dados “dataset\_netflix” para a resolução desse exercício.
- Utilize Python, R, SQL ou Scala para a resolução do exercício.
- Entregue seus códigos em um notebook de sua preferência ou sinta-se à vontade para duplicar os slides abaixo para responder.
- O problema consiste em montar um modelo de machine learning que preveja qual nota um filme receberia caso fosse colocado no catálogo. Não se preocupe muito com a precisão final do modelo. Os itens que serão avaliados individualmente são os seguintes:
  - 1) Tratamento dos dados.
  - 2) Feature engineering.
  - 3) Divisão da base de dados entre dataset teste e dataset treino.
  - 4) A Matriz de Confusão do seu modelo de testes assim como o gráfico de precision e recall do seu modelo
  - 5) Tendo em vista o resultado final o que você faria para melhorar o modelo?



# RESOLUÇÃO DO DESAFIO

Utilize esse espaço para resolver o desafio.





