***TESTE 1 -***

1. Exercício 1

Suponha que você possui uma base de dados rotulada com 10 classes não balanceadas, essa base é formada por 40 features de metadados e mais 3 de dados textuais abertos.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

* 1. *Descreva como faria a modelagem dessas classes.*

*Seguiria a metodologia padrão de projetos de ciência de dados:*

* + 1. *Entendimento e Análise Exploratória: AED das variáveis metadados, para analisar os comportamentos e padrões para cada classe;*
    2. *Pré-processamento: Normalização de dados numéricos, tratamento de valores ausentes, vetorização e processamento para as features textuais, como: TFIDF, COUNT VECT, WORD2VEC etc. Se for o caso, também realizar uma AED dessas variáveis textuais;*
    3. *Modelo: Experimentação com diferentes algoritmos de classificação, como Random Forest, SVM, Redes Neurais, entre outros, para encontrar o modelo mais adequado para os dados.*
    4. *Seleção de Features: Após treinar o modelo, Identificação das features mais relevantes para as classes através de métodos de seleção, como análise de importância.*
    5. *Avaliação do Modelo: Utilização de métricas de validação como: precisão, recall, F1-score e validação cruzada para avaliar a performance do modelo.*
    6. *Tuning de Hiperparâmetros: Se for o caso, aplicar um ajuste dos hiperparâmetros do modelo escolhido para otimização de desempenho.*

*Para essas atividades seria utilizado o Python com as respectivas bibliotecas: pandas, numpy, nltk (gensim ou spacy), matplotlib/seaborn, sklearn (pytorch).*

* 1. *Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?*
     1. *Resumo do processo de modelagem, desde a AED até a escolha do modelo final.*
     2. *Métricas de desempenho do modelo e como ele se comporta em relação aos dados disponíveis.*
     3. *Importância das features na classificação, se utilizado.*
     4. *Possíveis limitações do modelo e recomendações para o uso.*

*Para apresentação poderia ser utilizado o powerpoint e alguns recursos da apresentação podem ser desenvolvidos com python com suas bibliotecas de gráficos como: matplotlib/seaborn*

* 1. *Como faria a validação desse modelo?*

*Por ser um modelo de classes desbalanceadas, utilizaria as métricas: precisão, recall, F1-score e validação cruzada.*

* 1. *Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?*

*Seria necessário desenvolver uma pipeline bem estruturada e automatizada para realizar todo processo de forma (semi)automatizada.*

*Neste caso poderia ser utilizado o python com as libs padrões, ou analisar a necessidade de usar um pyspark para evitar possíveis problemas de performance com o pandas.*

* 1. *Como levaria esse projeto para um ambiente produtivo?*

*Poderia ser desenvolvida uma API com django, flask ou FastAPI, ou uma virtualização utilizando recursos como Docker.*

*EXTRA - Existe mais algo que gostaria de relatar sobre esse caso?*

1. Exercício 2:

Suponha que você tenha uma base de dados de vendas de uma loja de varejo que inclui informações sobre produtos, clientes, datas de compra e valores das vendas. A base de dados possui, em média, 10.000 registros diários.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

* 1. Como você iria explorar os dados para obter insights sobre o desempenho das vendas.

Iria fazer algumas análises descritivas e de séries temporais, assim, podendo identificar alguns pontos importantes como, distribuição dos valores de vendas, quantidade vendidas, comportamento das vendas ao longo do tempo, identificar padrões nos produtos vendidos, analisar dados dos clientes dentre outras características que fossem importantes.

* 1. Como você responderia as seguintes questões:
  2. Qual é o desempenho de vendas ao longo do tempo?

Poderia desenvolver algumas visualizações com vendas diárias, semanais, mensais, trimestrais etc para observar padrões, tendências temporais

* 1. Quais são os produtos mais vendidos?

Poderia calcular a frequência absoluta/relativa dos produtos por períodos como diárias, semanais, mensais.

* 1. Como as vendas variam por categoria de produtos?

Poderia calcular a frequência absoluta/relativa das categorias ao longo do tempo, como diárias, semanais, mensais.

* 1. Qual é a distribuição dos valores de venda?

Poderia utilizar visualizações como histograma ou boxplot.

* 1. Como os preços dos produtos afetam as vendas?

Para este caso é necessário analisar a relação do preço x quantidade de vendas. No caso ter uma tabela com o valor de cada produto e os respectivos volumes de venda. A partir dai podemos analisar via gráfica de dispersão ou com cálculo de correlação, como o coeficiente de pearson.

* 1. Qual é o perfil dos principais clientes em termos de compras?

Pode analisar a frequência e recência de compras, valor médio das compras, produtos/categorias mais compradas, períodos que costumam comprar dentre outras informações úteis.

* 1. Como você faria para identificar grupos de clientes nessa base de dados?

Utilizando técnicas de cluster como K-Means ou DBSCAN. Uma outra opção que está ficando popular na área é o algoritmo RFM (Recency, frequency, monetary), uma opção a ser avaliada.

* 1. Qual teste estatístico você usaria para provar uma hipótese referente aos segmentos de clientes? e como iria aplicá-lo?

Usaria Teste ANOVA ou Teste t: para comparar médias de compras ou valor de compras entre os segmentos de clientes

Ou, Teste Qui-quadrado para comparar distribuições categóricas entre os grupos, como: frequência de produtos comprados.

Extra - Pensando nos dados acima, seria possível fazer mais algum tipo de análise?

Seria possível desenvolver um algoritmo para fazer projeção de vendas ou de valor de vendas, o que poderia ainda ajudar a ter um estoque mais assertivo para não faltar e não ficar com produtos parados, ou até mesmo produtos vencidos se for o caso.

Para essas atividades seria utilizado o Python com as respectivas bibliotecas: pandas, numpy, matplotlib/seaborn, sklearn (prophet).

1. Exercício 3

Suponha que você tenha uma base de dados contendo textos jurídicos, como decisões judiciais, petições e documentos legais. A base de dados inclui informações sobre o conteúdo do texto, data, jurisdição e outras informações relevantes. Seu objetivo é criar um sistema de recomendação que sugira textos jurídicos semelhantes a um texto de referência.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

* 1. *Descreva como você desenvolveria o sistema de recomendação que recebe um texto de referência e sugere os textos mais semelhantes a ele na base de dados.*

*Para desenvolver um modelo de recomendação de texto realizaria as etapas comuns de um projeto de NLP como, processamento/limpeza dos textos e transformação dos textos para alguma técnica de vetorização como word embedding, term frequency, TF-IDF dentre outras.*

*Após a preparação dos dados, poderíamos testar algoritmos de deep learning para gerar novos textos.*

* 1. *Como você avaliaria esse sistema de recomendação?*

*Para avaliar este sistema, temos algumas opções como:*

* *cálculos de similaridade entre o texto gerado e o texto de entrada*
* *avaliação do próprio usuário, indicar se a recomendação foi boa ou ruim*
* *testes A/B, caso haja mais de um modelo em teste*

1. *Suponha que novos textos jurídicos sejam adicionados diariamente. Como você manteria o sistema de recomendação atualizado e garantiria que ele continue a fornece recomendações relevantes?*

*Primeiro seria importante avaliar se esses novos textos se diferem tanto dos textos já utilizados no treinamento do modelo, assim, evitando retreinar o modelo desnecessariamente.*

*Assim, sendo necessário retreinar modelo temos algumas estratégias como:*

* *agendar retreinamento do modelo completo semanalmente ou quinzenalmente, mas vai depender se o treinamento do modelo é muito demorado ou não*
* *para casos de treinamentos muito demorados, pode-se utilizar a estratégia de tuning, ou seja, utilizar apenas esses novos dados para retreinar o modelo, o que consome bem menos tempo*

*Para estas demandas as ferramentas/bibliotecas utilizadas seriam: python, pandas/pyspark, nltk/gensim, sklearn/pytorch.*

***TESTE 2 –***

1. Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?

Técnicas para avaliar se há evidências suficientes para determinada hipótese se confirma ou não.

Existem testes para avaliar média, distribuições etc.

1. O que são redes generativas adversárias (GANs) e quais são os possíveis usos dessas redes?

É um tipo de algoritmo que é dividido em 2 etapas:

* gerador: que cria dados e insere na rede
* discriminador que recebe esses dados e tenta distinguir dos dados reais

Esse tipo de rede pode ser utilizado em geração de imagens, síntese de dados dentre outras aplicações.

1. O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?

Modelos de linguagem são algoritmos que são treinados para gerar texto, traduzir idiomas, escrever diferentes tipos de conteúdo criativo e responder a perguntas de forma interativa.

Os modelos de linguagem tradicionais são treinados em conjuntos de dados de texto geralmente de alguns assuntos específicos, podendo utilizar técnicas como redes neurais recorrentes (RNNs) para aprender a gerar texto que seja similar ao texto que foi usado para treiná-los.

Já as LLMs são um tipo de modelo de linguagem que é treinado em conjuntos de dados de texto muito maiores que as tradicionais, contendo diferentes temas/assunto.

O que difere um modelo de LLMs de um de linguagem tradicional é basicamente o tamanho do conjunto de dados de treinamento e técnica de transformadores dos textos para treinar os algoritmos.

1. Suponha que você tenha um conjunto de dados com três ou mais grupos para comparar e deseja determinar se há diferenças significativas entre eles. Descreva como você escolheria entre o teste ou outras técnicas estatísticas

Para comprar mais de 2 grupos é possível utilizar ANOVA para verificar se há diferenças significativas nas médias e as suposições são atendidas.

1. Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?

É importante para deixar os dados com mais qualidade possível e padronizando seus registros.

As etapas mais comuns são:

- converter tudo para minúsculo

- remoção de caracteres especiais

- remoção de pontuações

- remoção de stopwords

- lemmatization e steaming

- vetorização

1. Descreva o processo de vetorização de texto e como modelos de linguagem como o Word2Vec ou o TF-IDF podem ser usados para representar palavras e documentos.

Ambos os processos transformam o texto em vetor numérico para que seja possível utilizar em modelo de ML, porém, essas técnicas possuem um resultado diferente.

Word2vec cria vetores de palavras para representar seu significado. Já o T-IDF cria uma matriz com valores de frequência relativa no documento das respectivos tokens.

1. O que é a análise de sentimento em NLP e quais são os principais métodos para realizar essa tarefa? Como você avaliaria a eficácia de um modelo de análise de sentimento?

É representação “sentimental”(polaridade emocional) do texto. Geralmente são classificados em positiva, neutra e negativa.

Os principais métodos para essa tarefa é utilizar algoritmos de classificação, porém, os textos tem que ser previamente classificados para então treinar um modelo.

Por ser um modelo de classificação, utilizaria as métricas comuns como precisão, recall e F1-score etc.

1. Qual é a diferença entre a classificação de texto e o agrupamento (clustering) de texto em NLP? Em que situações cada um é mais apropriado?

A classificação atribui rótulos pré-definidos a textos, enquanto o agrupamento cria grupos sem rótulos com base na similaridade.

Para classificação é útil para tarefas como análise de sentimento, já o agrupamento é útil para segmentar textos.

1. Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.

NER é uma técnica para Identificação e classificação de entidades, como nomes de pessoas, organizações, locais, data dentre outros em um texto.

É utilizado quando se deseja extrair informações de documentos, resumos automatizados dentre outros.

1. Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

Existem diferentes formas de lidar com este problema, como:

* trabalhar com Subamostragem ou sobreamostragem (oversampling ou undersampling)
* geração de dados sintéticos
* uso de pesos nas classes para classes minoritárias com algoritmos específicos.

Existem alguns trabalhos/pesquisas para testar essas diferentes técnicas para lidar com este problema e demonstram que usar algoritmos específicos que consigam dar pesos nas classes são mais eficientes do que trabalhar com sub/sobreamostragem.

Outra forma é trabalhar com dados sintéticos, porém, além de ser custoso computacionalmente, e às vezes financeiramente, a opção anterior continua sendo a opção com melhor custo x benefício.

*TESTE 3 - CASE*

Amostra:



Contextualização:

O Base de dados canada\_amostra em formato CSV representa um conjunto de empresas do Canadá com a respectiva descrição de seus produtos, dados econômicos e localização.

Assim, podemos caracterizar cada variável:

name: nome da empresa;

description: descrição do produto da empresa;

employees: número de empregados da empresa;

total\_funding: Total de investimento já recebido pela empresa;

city: cidade;

subcountry: estado;

lat: latitude da cidade;

lng: Longitude da cidade.

1. Problema:

Deseja-se prospectar empresas que possuam soluções em \*\*tratamento de água\*\* , principalmente, elativas à : \*\*solutions on waste and water, Improve water quality and water efficiency use, water contamination, water for human consumption, water resources\*\* .

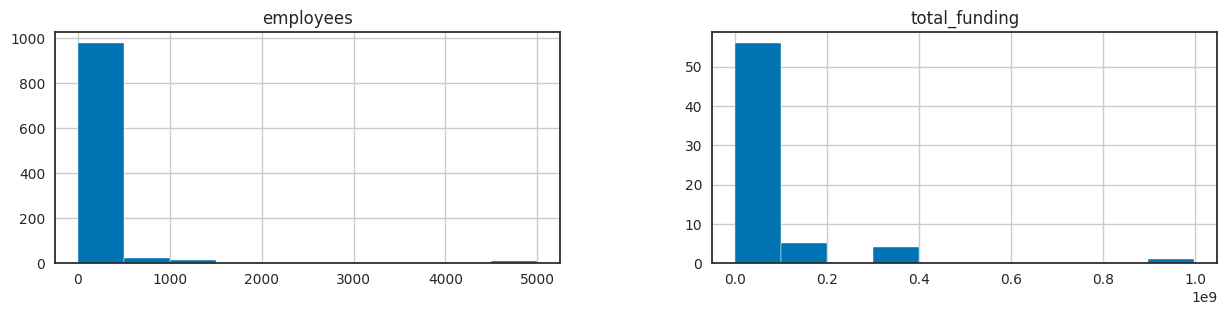
* 1. EXERCÍCIO 1 - Aplique um algoritmo de ML (ou um conjunto deles) capaz de selecionar as principais empresas indicadas para desenvolver a solução de acordo com seu alinhamento com o tema (Justifique a escolha do algoritmo).

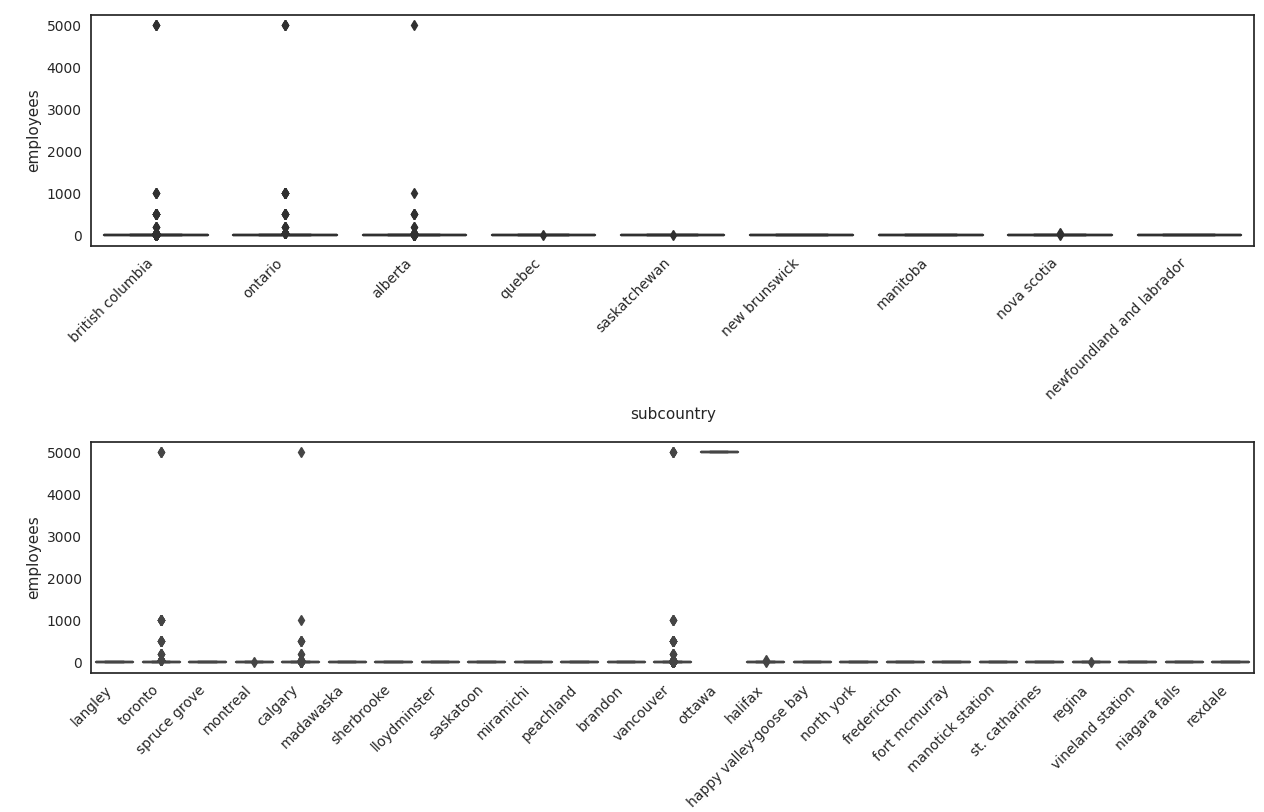
Inicialmente apliquei um algoritmo de LDA (Latent Dirichlet Allocation), uma solução popular que possibilita fazer uma modelagem de tópicos para cada documento, assim, possibilitando identificar de forma fácil os principais tópicos em cada texto. Além disso, é um modelo que performa bem, não necessitando de grandes recursos computacionais.

Com este resultado é possível identificar os principais tópicos em cada descrição das empresas.

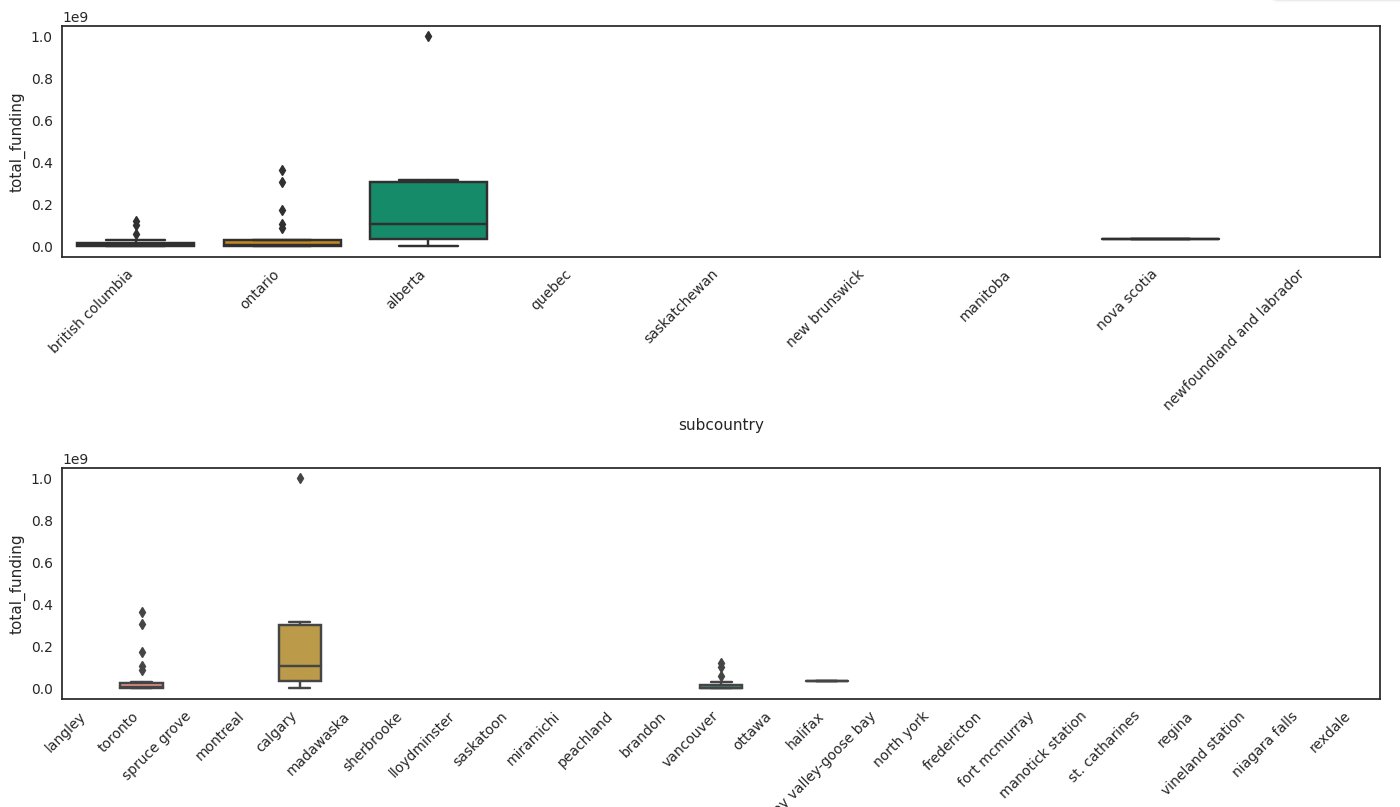
Outra opção que pode ser aplicado, individualmente ou em conjunto com LDA, algoritmos de cluster, DBSCAN, K-MEANS dentre outros. Estes algoritmos vai identificar a similaridade das descrições das empresas, agrupando palavras com mais proximidade, e também são recursos que não exigem alto poder computacional e fácil de analisar os resultados.(*esta opção não foi implementada por falta de tempo para finalizar o teste)*

* 1. EXERCÍCIO 2 - Faça uma análise exploratória dos resultados acrescentando as demais variáveis contidas no dataset. Quais insights você pode obter a partir desses dados? Quais são as principais cidades (pólos de desenvolvimento) para essa solução?





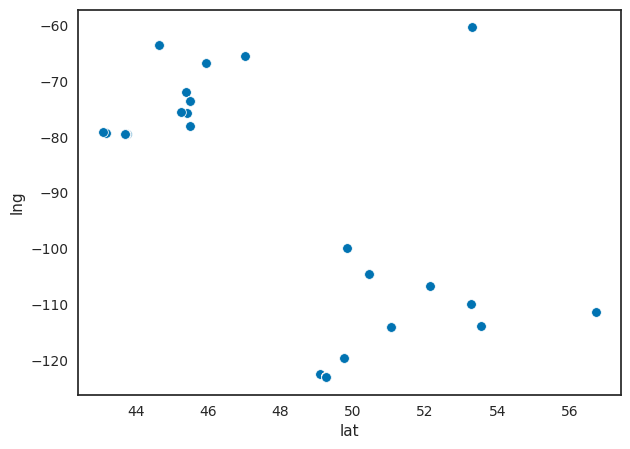
* Análises sobre empregados(employees):
  + Ao realizar as análises identifiquei que existem algumas empresas com 0 funcionários, o que pode ser considerado estranho ou o responsável pelas respectivas empresas não são considerados como funcionários.
  + também foi observado que a maioria das empresas possuem até 500 funcionários, e mais do que isso já são casos mais raros/outlier
  + e algumas cidades (city/subcounty) possuem a maior concentração de funcionários



* análises sobre investimentos (total\_funding):
  + foi identificado que a maioria das empresas tiveram investimento de até 30 milhões, e há algumas com investimentos maiores, mas são minorias e alguns outlier com investimentos acima de 100 milhões
  + como na quantidade de funcionários, os investimentos também estão concentrados em algumas cidades/subcountry.



* As principais cidades são: Vancouver(British Columbia), Toronto(Ontario) e Calgary(Alberta)
  + essas mesmas cidades são as que mais tem investimentos, também.



* Pela geolocalização podemos observar que as empresas estão mais concentradas em 2 regiões do Canadá.
  1. EXERCÍCIO 3 - EXTRA - Se você terminou o desafio de forma rápida, temos mais algumas perguntas para serem respondidas. Elas, como dito, não são obrigatórias, então sinta-se à vontade em não as responder ou até mesmo respondê-las parcialmente. Essa parte visa observar seu entendimento de um ambiente real de produção.
     1. a) organize seus códigos em pacotes garantindo seu versionamento e documentação (bibliotecas auxiliares, etc.).
     2. b) construa testes automatizados para validação do seu pacote.
     3. c) crie uma imagem Docker capaz de executar suas análises em um ambiente de produção.
     4. d) crie um GitHub público e suba todo o código do Teste 3, e disponibilize para avaliação.