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.
  2. Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?
  3. Como faria a validação desse modelo?
  4. Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?
  5. Como levaria esse projeto para um ambiente produtivo?

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.
  2. Como você responderia as seguintes questões:
  3. Qual é o desempenho de vendas ao longo do tempo?
  4. Quais são os produtos mais vendidos?
  5. Como as vendas variam por categoria de produtos?
  6. Qual é a distribuição dos valores de venda?
  7. Como os preços dos produtos afetam as vendas?
  8. Qual é o perfil dos principais clientes em termos de compras?
  9. Como você faria para identificar grupos de clientes nessa base de dados?
  10. Qual teste estatístico você usaria para provar uma hipótese referente aos segmentos de clientes? e como iria aplicá-lo?

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

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.
  2. Como você avaliaria esse sistema de recomendação?

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?

TESTE 2 –

1. Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?
2. O que são redes generativas adversárias (GANs) e quais são os possíveis usos dessas redes?
3. O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?
4. 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
5. Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?
6. 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.
7. 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?
8. 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?
9. Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.
10. Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

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).
  2. 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?
  3. 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.