

# Conjunto Questões

Atenção:

Sua resposta será avaliada com base na clareza, na profundidade conceitual, no raciocínio crítico e, quando for o caso, na qualidade dos exemplos ilustrativos. Os candidatos devem demonstrar uma sólida compreensão dos tópicos abordados nas questões.

## Questão 1

A representação computacional do texto é um dos problemas centrais em Processamento de Linguagem Natural (PLN), influenciando diretamente o desempenho, a interpretabilidade e a aplicabilidade dos modelos. Elabore um texto dissertativo comparando as seguintes técnicas de representação textual: modelos baseados em contagem (exemplos: Bag-of-Words, TF-IDF, entre outros), modelos probabilísticos (exemplos: n-gramas, modelos de linguagem clássicos, entre outros) e embeddings (exemplos: Word2Vec, GloVe, BERT, entre outros). Em sua resposta, você deve explicar os fundamentos conceituais de cada técnica, discutir as suas principais vantagens e limitações e analisar em quais tipos de aplicações cada abordagem tende a ser mais adequada.

## Questão 2

Imagine que uma empresa global de tecnologia está desenvolvendo um sistema de recomendação baseado em inteligência artificial para sugerir conteúdos personalizados a milhões de usuários. Durante os testes, foram identificados três problemas:

1. O modelo apresenta vieses que favorecem determinados grupos sociais.
2. Há dificuldade em garantir a reprodutibilidade dos resultados em diferentes ambientes.
3. A explicabilidade do modelo é limitada, o que dificulta a compreensão das decisões.

Além disso, a empresa precisa assegurar a responsabilidade no uso dos dados e a conformidade simultânea com o GDPR (Europa) e a LGPD (Brasil).

Pergunta: Como você estruturaria uma estratégia de governança ética para esse sistema, conciliando a inovação tecnológica com a responsabilidade social e legal? Em sua resposta, discuta os mecanismos para mitigação de vieses; as práticas que favoreçam a reprodutibilidade e a explicabilidade; as responsabilidades dos cientistas de dados e gestores e as medidas de conformidade com a GDPR e a LGPD. Defina os conceitos utilizados; apresente, de forma clara, os riscos e benefícios do uso combinado dos conceitos, e mostre algum exemplo que permita observar a relação da proposta de solução com o ciclo de operação sistêmica.

### Questão 3

Uma instituição pública deseja identificar fatores associados à evasão escolar utilizando um conjunto de dados integrado a partir de diferentes fontes administrativas (histórico acadêmico, indicadores socioeconômicos, registros de frequência e observações textuais de relatórios pedagógicos, entre outros). O banco resultante contém centenas de atributos, com diferentes escalas, níveis de qualidade e graus de redundância. A equipe pretende utilizar técnicas de mineração de dados para extrair conhecimento que possa subsidiar políticas públicas. Discuta, de forma conceitual e articulada, como as etapas de pré-processamento e seleção de atributos influenciam o processo de descoberta de conhecimento nesse cenário. Em sua resposta, espera-se que você analise criticamente o papel do pré-processamento na validade dos padrões identificados, discuta como a seleção de atributos pode afetar interpretabilidade, desempenho preditivo e estabilidade dos resultados, especialmente em contextos de alta dimensionalidade e possíveis correlações entre variáveis e explique por que a descoberta de conhecimento não se reduz à identificação de associações estatísticas, discutindo o risco de padrões espúrios, viés de seleção e problemas de generalização.