CONCURSO PÚBLICO PARA PROVIMENTO DE CARGOS DA CARREIRA DO MAGISTÉRIO SUPERIOR

Espelho resposta - Prova escrita

Em conformidade com as diretrizes de avaliação (referência ao item 14.3 do edital), este documento estabelece os tópicos e conceitos técnico-científicos considerados indispensáveis pela Comissão Examinadora para a obtenção da nota integral no tema sorteado para a prova escrita.

Tópico sorteado:

6. Inteligência Artificial Explicável: Interpretabilidade e explanabilidade (SHAP, LIME, Mapa de Saliencia, CAV, Attribution metrics). Modelos interpretáveis vs. modelos caixa-preta. Explicabilidade pré-modelo. Técnicas de visualização de modelos. Métodos Agnósticos.

A avaliação do candidato deve observar a completude, a profundidade teórica e a coerência na articulação dos itens abaixo.

1. Fundamentação

O candidato deve contextualizar a importância (ética, legal e crítica) do entendimento pelos humanos das decisões tomadas por modelos de Inteligência Artificial. Deve distinguir modelos interpretáveis e modelos caixa-preta.

Deve distinguir claramente *Interpretabilidade* (propriedade intrínseca do modelo de ser compreensível para humanos, ex: Regressão Linear, Árvores de Decisão) de *Explanabilidade/Explicabilidade* (técnicas ativas *post-hoc* aplicadas para elucidar as decisões de um modelo).

Discutir o contraponto entre desempenho preditivo e interpretabilidade, justificando por que modelos de Deep Learning, Random Forest ou SVMs com *kernels* não-lineares são considerados "caixa-preta" (alta complexidade, dimensionalidade latente e não-linearidade).

2. Explicabilidade Pré-Modelo

O candidato deve conceituar a explicabilidade pré-modelo (ou explicabilidade de dados), explicar que ela ocorre antes do treinamento e busca transparência e justiça (*fairness*) na base de dados.

O candidato deve mencionar e descrever técnicas e práticas aplicadas na explicabilidade pré-modelo, como por exemplo a Análise Exploratória de Dados (EDA) focada em identificação de vieses de representação, análise de distribuições e qualidade dos dados.

3. Métodos Agnósticos (LIME e SHAP)

O candidato deve apresentar a definição de Métodos Agnósticos, explicar que são técnicas *post-hoc,* que não dependem da arquitetura interna do modelo, tratando-o como uma caixa-preta (analisam apenas *inputs* e *outputs*).

O candidato deve explicar o mecanismo do LIME: a geração de instâncias perturbadas ao redor do dado original e o treinamento de um modelo substituto local interpretável (como regressão linear), ponderado pela proximidade.

O candidato deve conectar o SHAP à Teoria dos Jogos Cooperativos (Valores de Shapley). Deve mencionar a propriedade matemática da Eficiência ou Aditividade Local (onde a soma das contribuições das *features* é igual à diferença entre a predição local e a média global).

Ambas as técnicas devem ser abordadas com um nível de aprofundamento consistente, apresentando não apenas o conceito, mas descrevendo o funcionamento, a análise e como elas são computadas.

4. XAI em Redes Neurais Convolucionais

Explicar o que são mapas de saliência, dizer que são métodos de atribuição de pixels baseados em gradientes, apresentar exemplos como *Grad-CAM, Vanilla Gradient, Integrated Gradients*. Destacar que eles destacam onde a rede focou na imagem de entrada. O candidato deve apontar a limitação semântica destas técnicas (que mostram "onde", mas não o "por quê"). É desejável que o candidato discorra com mais profundidade sobre pelo menos uma das técnicas de mapas de saliência.

O candidato deve diferenciar o CAV (*Concept Activation Vectors*) dos mapas de saliência. O candidato deve explicar que o CAV trabalha com abstrações de alto nível, utilizando derivadas direcionais no espaço latente para medir a importância de conceitos definidos por humanos (ex: "listras", "textura") para a predição de uma classe.

5. Técnicas de Visualização de Modelos

O candidato deve mencionar técnicas de redução de dimensionalidade (como t-SNE ou UMAP) para projetar e visualizar *embeddings* de camadas ocultas, verificando a separabilidade linear e o agrupamento semântico das classes antes da decisão final.

Citar ferramentas como PDP (*Partial Dependence Plots*) para visualizar o efeito marginal global de uma ou duas variáveis na predição, ou citar a técnica ICE (*Individual Conditional Expectation*) para desagregação local.

6. Avaliação de Explicabilidade (*Attribution Metrics*)

O candidato deve deixar claro que gerar um mapa de calor ou vetor SHAP não garante que a explicação seja verdadeira em relação ao que o modelo de fato aprendeu. Destacar que explicabilidade não imputa causalidade.

O candidato deve citar Métricas de Fidelidade (*Faithfulness*), abordagens experimentais como Perturbação de *Features* (ex: *Deletion/Insertion AUC*, remover as características apontadas como mais importantes e observar a queda na acurácia do modelo) e/ou *Sanity Checks* (verificar se a explicação falha quando os pesos do modelo são aleatorizados).

7. Critérios transversais

Para fins de atribuição de nota, a banca observou:

- A precisão conceitual e rigor matemático/computacional na descrição dos métodos (especialmente SHAP, LIME, mapas de saliência e CAV).
- A habilidade do candidato em cruzar as informações, não apenas listando os métodos, mas contrapondo suas vantagens e limitações (ex: SHAP vs. LIME, Mapas de Saliência vs. CAV).
- A estruturação lógica do texto, uso de terminologia técnica adequada ao estado da arte e clareza na exposição do raciocínio.
- O poder de sistematização do conteúdo apresentado pelo candidato, bem como a clareza nos argumentos.
- A capacidade do candidato de sustentar e expandir os argumentos apresentados no texto perante os questionamentos da banca na etapa de arguição.
- A habilidade para responder prontamente a perguntas, de forma clara e objetiva, explicando conceitos complexos de forma intuitiva, sem depender exclusivamente da leitura do texto.