TECH CHALLENGE

FASE 1

TECH CHALLENGE

Tech Challenge é o projeto que engloba os conhecimentos obtidos em todas as disciplinas da fase. Esta é uma atividade que, em princípio, deve ser desenvolvida em grupo. É importante atentar-se ao prazo de entrega, uma vez que essa atividade é obrigatória, valendo 90% da nota de todas as disciplinas da fase.

Desafio

Um grande hospital universitário busca implementar um sistema inteligente de suporte ao diagnóstico, capaz de ajudar médicos e equipes clínicas na análise inicial de exames e no processamento de dados médicos.

Com um volume crescente de pacientes e exames, como radiografias, tomografias, ressonâncias e prontuários digitalizados, o hospital precisa de soluções que acelerem a triagem e apoiem as decisões médicas, reduzindo erros e otimizando o tempo dos profissionais.

Nesta primeira fase, o desafio é criar a base do sistema de IA focado em **machine learning**, permitindo que resultados de exames sejam analisados automaticamente e destacando informações relevantes para o diagnóstico.

Objetivo

Construir uma solução inicial com foco em IA para processamento de exames médicos e documentos clínicos, aplicando fundamentos essenciais de IA, Machine Learning e Visão Computacional.

Entregas técnicas

Processamento de dados médicos

- Classificação de exames com **Machine Learning**: você deve escolher uma base de dados em forma de **tabela** e realizar o diagnóstico: “a pessoa tem ou não uma doença”. Isso acontecerá via algoritmos de aprendizado de máquina.

- **EXTRA:** além do algoritmo de diagnóstico com dados estruturados, você **também** pode optar por realizar um diagnóstico com dados de imagem, utilizando **redes neurais convolucionais (CNN).** (**Observação:** isso não é obrigatório, mas pode aumentar a sua nota caso você não atinja a pontuação máxima na atividade).

Dados e Modelos

- Escolha um ou mais datasets médicos públicos e discuta o problema a ser resolvido.

Exploração de dados:

- Carregue a base de dados e explore suas características;
- Analise estatísticas descritivas e visualize distribuições relevantes, discutindo os resultados.

Pré-processamento de dados:

- Realize a limpeza dos dados, tratando valores ausentes e inconsistentes (se necessário);
- Pipeline de pré-processamento de dados em Python.
    - Converta variáveis categóricas e numéricas em formatos adequados para modelagem.
- Realize a análise de correlação.

Modelagem:

- Crie modelos preditivos de classificação utilizando duas ou mais técnicas à sua escolha (por exemplo: Regressão logística, Árvore de Decisão, KNN etc);
- Separação clara entre treino, validação e teste.

Treinamento e avaliação do modelo:

- Treine o modelo com o conjunto de treinamento;

- Avaliação do modelo com os dados de teste e métricas adequadas (accuracy, recall, F1-score). Discuta a escolha da métrica considerando o problema;
- Apresente uma interpretação dos resultados (utilize técnicas como feature importance e SHAP);
- Discuta os resultados de maneira crítica. O seu modelo pode ser utilizado na prática? Como? (Lembre-se que o médico sempre deve ter a palavra final no diagnóstico).

Exemplo de fontes de dados que podem ser utilizadas neste desafio:

- Tarefa principal a ser avaliada:
  - Diagnóstico de câncer de mama (maligno ou benigno): https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data;
  - Diagnóstico de diabetes: https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data
  - Outro de sua preferência.
- Tarefa extra - visão computacional:
  - Detecção de Pneumonia em Radiografias: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
  - Detecção de câncer de mama: https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset/data

Código e Organização

- Projeto em Python estruturado e documentado;
- Notebook Jupyter ou scripts Python para demonstração dos resultados.

Entregáveis da Fase 1

Arquivo PDF com link do repositório Git:

- Código-fonte completo;

- Dockerfile e README.md com instruções de execução;
- Dataset (ou link para download);
- Resultados obtidos (prints, gráficos e análises);
- Relatório técnico explicando:
    - Estratégias de pré-processamento;
    - Modelos usados e por quê;
    - Resultados e interpretação dos dados.

Vídeo de demonstração:

- Realizar o upload no YouTube ou Vimeo (usando as configurações “público” ou “não listado”). A duração do vídeo deve ser de até 15 minutos;
- Demonstração do sistema em execução com breve explicação do fluxo.


POS TECH