

POSTECH

TECH CHALLENGE

FASE 1

TECH CHALLENGE

Tech Challenge é o projeto que engloba os conhecimentos obtidos em todas as disciplinas dessa 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

Uma rede de hospitais e centros de saúde especializados no atendimento à mulher busca implementar um sistema inteligente de suporte ao diagnóstico e detecção de riscos, capaz de ajudar profissionais de saúde na identificação precoce de condições que afetam a segurança e saúde feminina.

Com um volume crescente de pacientes e a necessidade de identificar rapidamente situações de risco, desde doenças como câncer de mama até sinais de violência doméstica em prontuários médicos, a instituição precisa de soluções que acelerem a triagem, apoiem as decisões médicas e garantam um atendimento mais seguro e eficaz para as mulheres.

Nesta primeira fase, o desafio é criar a base do sistema de IA focado em Machine Learning, permitindo que dados médicos sejam analisados automaticamente para identificar padrões de risco relacionados à segurança e saúde das mulheres.

Objetivo

Construir uma solução inicial com foco em IA para processamento de dados médicos relacionados à saúde e segurança da mulher, 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 relacionada à saúde feminina e realizar a classificação de riscos ou diagnósticos via algoritmos de aprendizado de máquina;
- EXTRA: além do algoritmo de diagnóstico com dados estruturados, você pode optar por realizar também um diagnóstico com dados de imagem (mamografias, ultrassons), utilizando redes neurais convolucionais (CNN). **Observação**: não é obrigatório, mas pode aumentar sua nota caso não atinja a pontuação máxima.

Dados e Modelos

- Escolha um ou mais datasets públicos relacionados à segurança e saúde da mulher 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;
- Identifique padrões específicos relacionados à saúde e segurança feminina.

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 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, considerando técnicas de explicabilidade da previsão (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 fonte de dados que pode ser utilizada neste desafio (sugestão)

- Tarefa principal a ser avaliada:
 - Diagnóstico de câncer de mama (maligno ou benigno):
<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data>
 - Detecção de síndrome dos ovários policísticos (SOP):
<https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>
 - Outro de sua preferência
- EXTRA - Visão computacional:
 - Detecção de câncer de mama em mamografias:
<https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset>

Código e Organização

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

Entregáveis da Fase 1

Um arquivo em PDF com:

Link do repositório Git

- Código-fonte completo;
- README.md com instruções de execução;
- Dockerfile (se usar docker);
- Dataset (ou link para download);
- Resultados obtidos (prints, gráficos e análises);
- Relatório técnico explicando:
 - Discussões da análise exploratória.
 - Estratégias de pré-processamento.
 - Modelos usados e porquê.
 - Resultados e interpretação dos dados.

Vídeo de demonstração

- Upload no YouTube ou Vimeo (público ou não listado) de até 15 minutos;
- Demonstração do sistema em execução com breve explicação do fluxo.



POSTECH