

---

## Tech Challenge Fase 1 - Análise Preditiva de Câncer de Pele com IA

**Aluno:** Fernando Stuque Alves




**RM:** 366142

**Pós-Graduação em IA para DEVs - 6AIDT**

**FIAP - Faculdade de Tecnologia**

---

### Links do Projeto:

-  **Repositório GitHub (Código-Fonte Completo):**  
[https://github.com/Ferstuque/skin\\_cancer\\_analysis](https://github.com/Ferstuque/skin_cancer_analysis)
  -  **Vídeo de Demonstração (YouTube):**  
<https://www.youtube.com/watch?v=20H0Xxup-zE>
  -  **Dataset HAM10000:**  
<https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000>
- 

### Resumo

Este relatório detalha o desenvolvimento de uma solução de Inteligência Artificial para o Tech Challenge da FIAP, focada em apoiar o diagnóstico de câncer de pele. Utilizando o dataset HAM10000, foram criados e avaliados dois modelos distintos: um modelo de Machine Learning (XGBoost) baseado em dados clínicos, e um modelo de Visão Computacional (CNN com ResNet50V2) para análise de imagens. O modelo XGBoost demonstrou performance superior na tarefa crítica de detecção de Melanoma, alcançando um Recall de 94%. A solução completa foi empacotada em uma aplicação web interativa com API (FastAPI) e Dashboard (Streamlit), utilizando Docker para garantir a portabilidade e reprodutibilidade. O projeto conclui que a abordagem mais eficaz combina a análise contextual dos dados do paciente com a análise visual das lesões.

---

## 1. Introdução

### 1.1. O Problema de Negócio

O desafio proposto simula uma necessidade de um grande hospital universitário: a criação de um sistema de IA para auxiliar dermatologistas na triagem e diagnóstico de lesões de pele. O objetivo é aumentar a precisão, reduzir o tempo de diagnóstico e otimizar o fluxo de trabalho dos especialistas, especialmente na detecção precoce do Melanoma, o tipo mais agressivo de câncer de pele.

### 1.2. O Dataset: HAM10000

Para este projeto, utilizamos o dataset (HAM10000). Esta base de dados é ideal por ser multimodal, contendo:

- **10.015 imagens** de lesões de pele, divididas em 7 categorias diagnósticas.
- Um arquivo de **metadados tabulares** (HAM10000\_metadata.csv) com informações clínicas para cada imagem, como idade, sexo e localização anatômica da lesão.

### 1.3. A Solução Proposta

A solução foi desenvolvida em duas frentes complementares:

1. **Modelo Preditivo Tabular:** Focado em classificar uma lesão como "Melanoma" ou "Não-Melanoma", utilizando as informações contextuais do paciente.
2. **Modelo de Visão Computacional:** Focado em classificar a imagem da lesão em uma das 7 categorias, utilizando apenas a informação visual.

Ambos os modelos foram integrados em uma aplicação web para demonstração e interação.

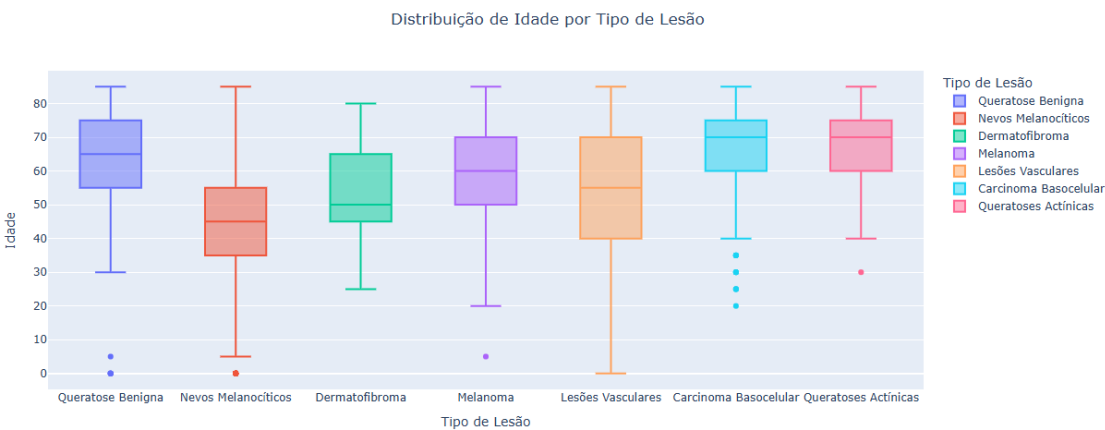
---

## 2. Análise Exploratória e Pré-processamento

### 2.1. Principais Insights da Análise Exploratória (EDA)

A análise inicial dos dados revelou pontos cruciais que guiaram todo o projeto:

- **Severo Desbalanceamento de Classes:** A classe "Nevo Melanocítico" (nv), majoritariamente benigna, compõe 67% do dataset, enquanto o "Melanoma" (mel) representa apenas 11%. Isso tornou a métrica de acurácia inadequada e direcionou o foco para o Recall.
- **Importância da Idade:** A análise demonstrou uma forte correlação entre a idade do paciente e o tipo de lesão, validando a idade como uma feature preditiva de alto valor.



2.2. Pipeline de Pré-processamento

Um script reutilizável (src/data\_preprocessing.py) foi criado para realizar o tratamento dos dados tabulares, incluindo:

- Preenchimento de valores ausentes na coluna age com a mediana.
- Aplicação de One-Hot Encoding em variáveis categóricas (sex, localization, dx\_type).
- Padronização dos nomes das colunas para garantir a compatibilidade.

3. Modelo Campeão: XGBoost para Dados Tabulares

3.1. Estratégia e Resultados

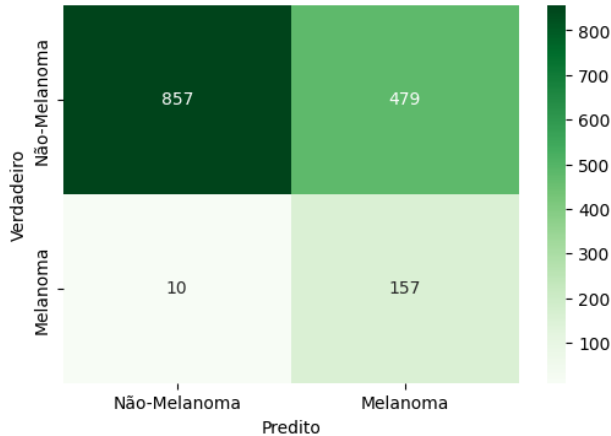
O objetivo principal era maximizar a detecção de Melanoma. Após um benchmark com 6 algoritmos, o **XGBoost** foi selecionado e otimizado com GridSearchCV. O modelo final, avaliado no conjunto de teste, alcançou resultados excelentes para o contexto clínico:

- **Recall (Sensibilidade): 94.0%**
- **Acurácia Balanceada: 85.3%**
- **AUC: 93.8%**

Um Recall de 94% significa que o modelo identificou corretamente 94 de cada 100 casos de Melanoma, cumprindo seu papel como uma ferramenta de segurança de alta eficácia.

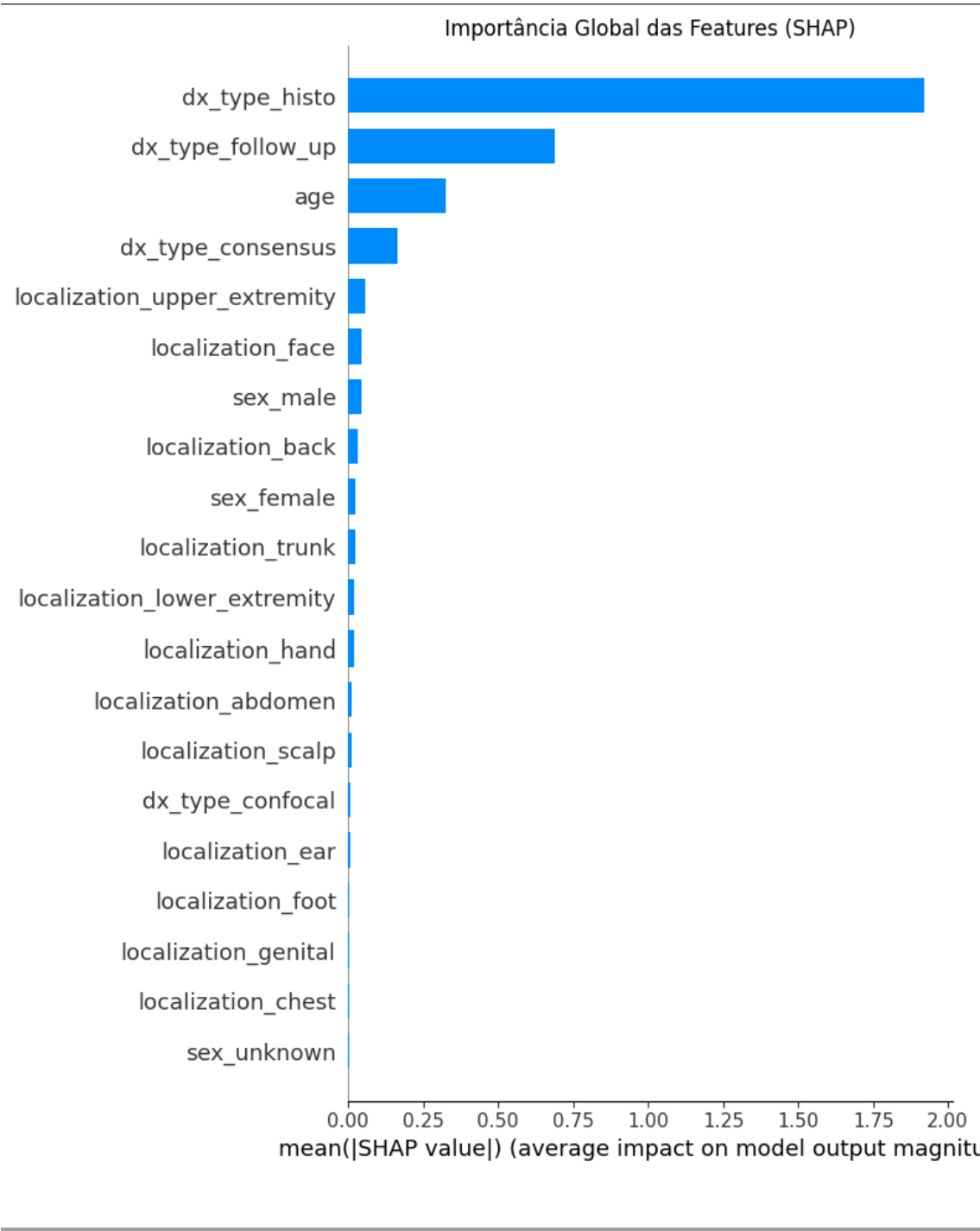
Métricas de Classificação		Relatório de Classificação Final				
Métrica	Valor	Classe	Precisão	Recall	F1-Score	Suporte
Acurácia	6.747	Não-Melanoma	0.99	0.64	0.78	1336
Precisão	2.469	Melanoma	0.25	0.94	0.39	167
Recall (Sensibilidade)	9.401	Acurácia			0.67	1503
F1-Score	3.910	Macro Avg	0.62	0.79	0.58	1503
AUC-ROC	8.362	Weighted Avg	0.91	0.67	0.74	1503

Matriz de Confusão Final - XGBoost Otimizado com GridSearchCV



3.2. Interpretabilidade com SHAP

Para garantir a transparência do modelo, utilizamos a biblioteca SHAP. A análise revelou que as decisões do modelo são impulsionadas por fatores clinicamente relevantes, sendo as mais importantes o **tipo de diagnóstico prévio** e a **idade do paciente**, o que confere alta confiabilidade à sua lógica interna.



4. Tarefa Extra: Visão Computacional com ResNet50V2

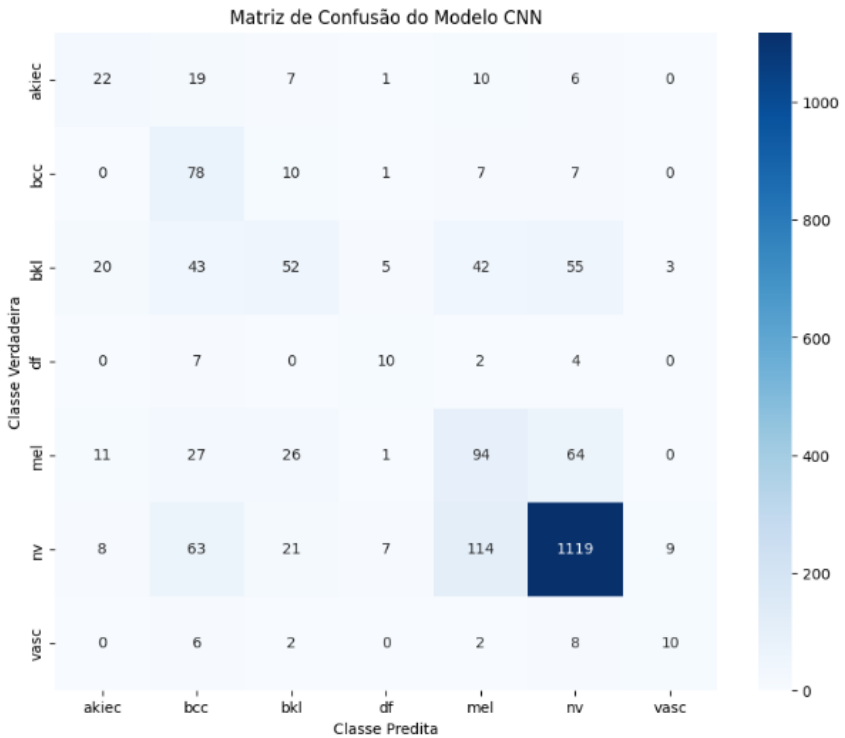
4.1. Estratégia e Resultados

Para a análise de imagens, implementamos uma CNN com a arquitetura ResNet50V2 via Transfer Learning. Utilizamos técnicas como `class_weight` e `fine-tuning` para combater o desbalanceamento. O modelo final alcançou uma **Acurácia Balanceada de 48.3%** e um **Recall de 42.2% para Melanoma**.

--- Métricas Gerais do Modelo CNN (ResNet - Fine-Tuned) ---  
Acurácia Geral: 0.6915  
Acurácia Balanceada: 0.4829  
AUC Ponderada (One-vs-One): 0.8837

--- Relatório de Classificação por Classe ---

	precision	recall	f1-score	support
akiec	0.360656	0.338462	0.349206	65.000000
bcc	0.320988	0.757282	0.450867	103.000000
bkl	0.440678	0.236364	0.307692	220.000000
df	0.400000	0.434783	0.416667	23.000000
mel	0.346863	0.421525	0.380567	223.000000
nv	0.885986	0.834452	0.859447	1341.000000
vasc	0.454545	0.357143	0.400000	28.000000
accuracy	0.691463	0.691463	0.691463	0.691463
macro avg	0.458531	0.482858	0.452064	2003.000000
weighted avg	0.719340	0.691463	0.696454	2003.000000



4.2. Análise Comparativa e Sinergia

A comparação entre os dois modelos é o insight mais valioso:

- O **XGBoost (tabular)** é superior na segurança da detecção de Melanoma (Recall 94%) por usar o **contexto clínico**.
- A **CNN (imagem)** oferece uma análise visual complementar, capaz de identificar 42% dos melanomas e se destacar em outras classes.

## 5. Arquitetura da Solução e Aplicação Final

A solução completa foi encapsulada em uma aplicação web interativa, garantindo a portabilidade com Docker. A arquitetura consiste em:

- **API Backend (FastAPI):** Serve as predições do modelo XGBoost.
- **Frontend (Streamlit):** Provê uma interface para o usuário interagir com ambos os modelos.
- **Containerization (Docker):** Empacota toda a aplicação e suas dependências.

**Previsão de Melanoma com Base em Dados Clínicos**

Preencha os campos abaixo com as informações da lesão para receber uma predição do modelo XGBoost, que é especializado em identificar o risco de Melanoma.

**Informações do Paciente**

Idade do Paciente: 75

Sexo: Feminino

Analisar Risco de Melanoma

**Características da Lesão**

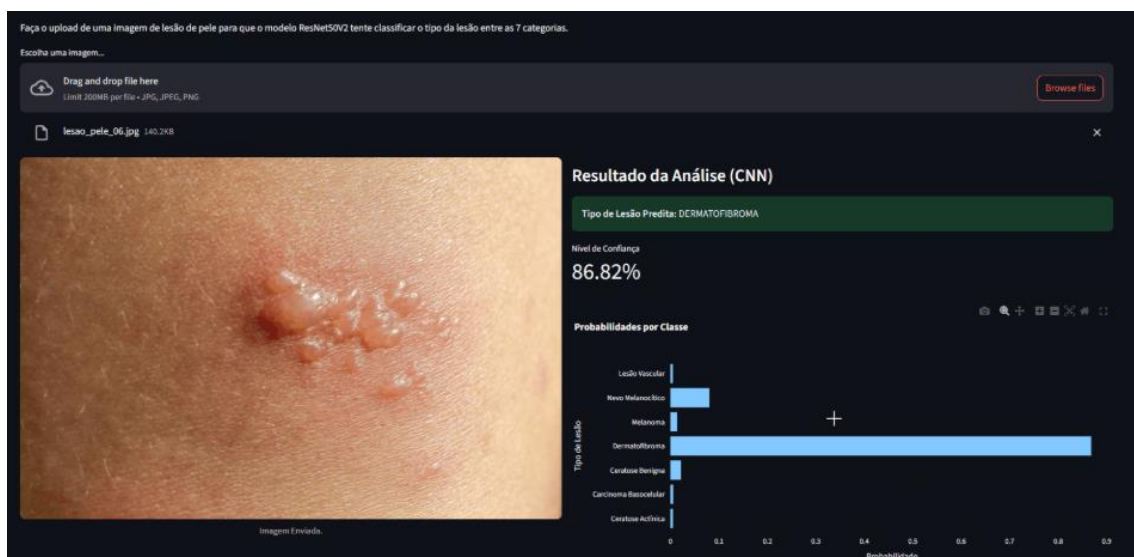
Localização da Lesão: orelha

Método de Confirmação Inicial: Acompanhamento

**Resultado da Análise (XGBoost)**

Diagnóstico Sugerido: Não-Melanoma

Probabilidade de ser Melanoma: 4.28%



## 6. Discussão Crítica

- **Uso na Prática:** A solução é proposta como um **sistema de apoio à decisão**, para auxiliar na triagem e priorização de casos, mas nunca para substituir o diagnóstico final do médico.
  - **Limitações:** O modelo foi treinado em um dataset público e pode conter vieses. Sua implementação em um cenário real exigiria validação com dados locais e aprovação de órgãos reguladores (ANVISA). A qualidade da imagem de entrada também é um fator crítico para o modelo CNN.
  - **O Papel do Especialista:** A palavra final é e sempre deve ser do profissional de saúde. A IA é uma ferramenta para aumentar a capacidade humana, não para substituí-la.
- 

## 7. Conclusão

Este projeto entregou com sucesso uma solução de IA ponta a ponta, desde a análise de dados até uma aplicação funcional e empacotada. Demonstrou-se que, para este problema, a combinação de dados clínicos e de imagem, com um foco maior no contexto do paciente, oferece o caminho mais promissor para a criação de ferramentas de diagnóstico eficazes e seguras.