**🔬 Inspiração Científica**

Este projeto foi **diretamente inspirado** na pesquisa:  
[**"Systems analysis reveals distinct molecular signatures associated with immune responsiveness to influenza vaccination"**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7062420/) (PMC7062420) publicado na *Nature Immunology*, adaptado para fins de aprendizado em machine learning.

**📌 Visão Geral do Projeto**

Modelo preditivo que identifica **high-responders** e **low-responders** à vacinação com base em expressão gênica, utilizando:

* **Random Forest** com otimização de hiperparâmetros
* **Balanceamento via class\_weight** (sem SMOTE)
* **Interpretação por SHAP values**
* **Validação cruzada estratificada**

**📊 Métricas de Desempenho**

| **Métrica** | **Validação Cruzada** |
| --- | --- |
| **Acurácia** | 78% |
| **F1-Score** | 79% |
| **Precision** | 81% |
| **Recall** | 78% |

**Melhores parâmetros:**

{

'class\_weight': 'balanced', # Balanceamento interno

'max\_depth': None,

'n\_estimators': 100

}

**🛠️ Estratégias de Balanceamento**

Diferentemente de abordagens com SMOTE, utilizamos:

1. **Pesos customizados no Random Forest**

python

Copy

RandomForestClassifier(class\_weight='balanced')

1. **Validação cruzada estratificada** para preservar proporções
2. **Ajuste de threshold** (não implementado, mas sugerido como melhoria futura)

**🚀 Próximos Passos**

* Testar XGBoost com scale\_pos\_weight
* Implementar análise de curva ROC
* Adicionar explicações individuais com SHAP

**📂 Estrutura do Projeto**

Copy

├── data/

│ └── simulated\_data.csv # Dados com padrões do artigo

├── notebooks/

│ └── Model\_Training.ipynb # Código completo

└── references/

└── PMC7062420.pdf # Artigo original

Este projeto demonstra:

* Como lidar com desbalanceamento **sem oversampling**
* A importância do class\_weight em Random Forest
* Interpretação de modelos para dados biológicos