

# Relatório Técnico: Análise Comparativa de *Pipelines* de *Machine Learning* para Diagnóstico de Diabetes

Adriel Felipe Cândido Santos - [12.adriel@gmail.com](mailto:12.adriel@gmail.com), João Marcos Simões - [joao\\_marcos99@hotmail.com](mailto:joao_marcos99@hotmail.com), Letícia Rodrigues Nepomucena Lopes - [lehnep2@gmail.com](mailto:lehnep2@gmail.com), Lucas Felipe Silva - [lucfsilva@gmail.com](mailto:lucfsilva@gmail.com), Vagner Barbosa Dantas - [contato@vagnerbarbosa.com](mailto:contato@vagnerbarbosa.com)

## 1.0 Introdução

Este relatório técnico é uma **entrega obrigatória da Fase 1 do Tech Challenge IADT**, que engloba os conhecimentos obtidos nas disciplinas dessa fase. O desafio central é apoiar um **grande hospital universitário** que busca implementar um **sistema inteligente de suporte ao diagnóstico**. Este sistema visa ajudar médicos e equipes clínicas na análise inicial de exames e dados, apoiando decisões médicas, **reduzindo erros e otimizando o tempo dos profissionais**. Nesta primeira fase, o foco é **criar a base do sistema de IA focado em machine learning**, realizando a classificação de exames com Machine Learning para diagnosticar se "a pessoa tem ou não uma doença".

O projeto utiliza o **Pima Indians Diabetes Dataset** e, para estabelecer essa fundação de ML, **compara duas abordagens distintas (Pipelines A e B)** de pré-processamento, modelagem e avaliação. Os resultados detalhados desta análise comparativa encontram-se no **Notebook** ([Diabetes\\_Analysis.ipynb](#)).

Este documento está estruturado para fornecer uma visão abrangente do projeto. Começa por delinear a metodologia central e o ambiente técnico, seguido por uma descrição detalhada das arquiteturas dos dois *pipelines*. Subsequentemente, o relatório apresenta a estrutura de avaliação, os resultados quantitativos da análise comparativa e uma discussão que interpreta esses resultados. O relatório conclui com uma recomendação final baseada na evidência empírica coletada. As seções a seguir detalham a metodologia, a arquitetura e a estrutura de avaliação que sustentam esta

análise.

---

## 2.0 Metodologia Central e Ambiente Técnico

Estabelecer uma base consistente e reproduzível é fundamental para qualquer experimento científico. Para garantir uma comparação justa e confiável entre os dois *pipelines*, decisões importantes relativas ao **conjunto de dados**, ao **ambiente técnico** e ao **tratamento proativo de desafios comuns de dados** foram padronizadas em todo o projeto.

### 2.1 Descrição do Conjunto de Dados

A análise foi conduzida utilizando o **Pima Indians Diabetes Dataset**, um conhecido conjunto de dados públicos para tarefas de classificação binária. Para garantir consistência e reprodutibilidade, o conjunto de dados foi obtido diretamente do Kaggle para ambos os *pipelines* usando a biblioteca kagglehub.

Fonte: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data>

### 2.2 Ambiente Técnico e Reprodutibilidade

Para garantir um ambiente isolado e livre de conflitos para a execução, o método recomendado é o **Docker**. Essa abordagem encapsula todas as dependências e garante que os resultados possam ser replicados de forma confiável. Para execução local alternativa, o seguinte ambiente é necessário:

1. **Versão do Python:** O projeto é compatível com Python **3.10 ou 3.11**.
2. **Bibliotecas Principais:** As dependências-chave, conforme listadas no arquivo requirements.txt, incluem pandas, numpy, scikit-learn, matplotlib, seaborn, imblearn, xgboost, kagglehub e shap.

### 2.3 Estratégia Comum de Pré-processamento: Desequilíbrio de Classes

Um potencial **desequilíbrio de classes** dentro do conjunto de dados foi identificado como um risco significativo para o desempenho do modelo, pois poderia levar a um

viés contra a classe minoritária (pacientes com diabetes). Para mitigar proativamente esse problema, a técnica **SMOTE** (*Synthetic Minority Over-sampling Technique*) foi aplicada como uma etapa padrão de pré-processamento em ambos os *pipelines*.

Com esta metodologia fundamental estabelecida, o relatório detalhou as diferenças arquitetônicas específicas entre os dois *pipelines* em comparação.

---

## 3.0 Arquiteturas dos *Pipelines*

O cerne desta análise envolve uma comparação direta de duas abordagens distintas para a tarefa de diagnóstico: uma linha de base direta projetada para **simplicidade e implantação rápida (Pipeline A)**, e uma abordagem mais complexa e otimizada, projetada para máxima **sensibilidade diagnóstica (Pipeline B)**. Esta seção relaciona os componentes e estratégias específicas que definem cada *pipeline*.

### 3.1 Pipeline A: A Abordagem Linha de Base/MLOps

O **Pipeline A** serve como o modelo de linha de base, representando uma abordagem padrão e robusta para um problema de classificação de *machine learning*. Sua arquitetura é caracterizada por sua simplicidade e confiança em técnicas bem estabelecidas.

- **Imputação:** Valores ausentes são tratados usando o **KNN Imputer**, que estima um valor com base nos "k-vizinhos mais próximos" no espaço de recursos (*feature space*).
- **Modelagem:** O *pipeline* emprega um conjunto de algoritmos de *machine learning* "clássicos" para classificação, fornecendo um sólido *benchmark* de desempenho.

### 3.2 Pipeline B: A Abordagem Otimizada/Deep Dive

O **Pipeline B** é um modelo aprimorado, especificamente projetado para melhorar a linha de base incorporando técnicas mais avançadas. Seu objetivo principal é aumentar a **sensibilidade diagnóstica**, um requisito crítico em um ambiente clínico.

- **Pré-processamento Avançado:** Este *pipeline* incorpora **Engenharia de Recursos** (*Feature Engineering*), criando novas variáveis de interação, como `idade_imc` e `glicose_imc`, para capturar relacionamentos complexos nos dados.

Inclui também uma etapa deliberada de **remoção de recursos** para reduzir o ruído, mitigar a multicolinearidade e melhorar a generalização do modelo.

- **Modelagem Avançada:** Além dos modelos clássicos, este *pipeline* inclui o algoritmo **XGBoost**, conhecido por seu alto desempenho em tarefas de classificação de dados estruturados.
- **Limiar Estratégico** (*Strategic Thresholding*): Um limiar de classificação de **0.3** é explicitamente aplicado. Essa decisão foi tomada para deliberadamente **priorizar e maximizar a métrica Recall**, tornando o modelo mais sensível à detecção de casos positivos de diabetes.

A seção a seguir detalha a estrutura utilizada para avaliar e comparar rigorosamente o desempenho destas duas arquiteturas distintas.

---

## 4.0 Estrutura de Avaliação

A seleção de métricas de avaliação apropriadas é de importância crítica em um contexto clínico. Para tarefas de diagnóstico, nem todos os erros de previsão têm o mesmo peso. A estrutura de avaliação deve, portanto, refletir a prioridade clínica de **minimizar diagnósticos perdidos**, onde a falha em identificar um paciente com uma condição pode ter consequências graves.

### 4.1 Seleção de Métricas de Desempenho

As seguintes métricas foram escolhidas para fornecer uma avaliação abrangente do desempenho do modelo, com um foco claro na **utilidade clínica**.

- **Recall (Sensibilidade):** Definido como a porcentagem de casos positivos reais (pacientes com diabetes) que foram corretamente identificados pelo modelo. Esta métrica é **essencial** para este caso de uso, pois o objetivo principal é **minimizar falsos negativos (FN)** e garantir que o maior número possível de casos verdadeiros seja detectado.
- **Precision (Precisão):** Definida como a porcentagem de previsões positivas que estavam, de fato, corretas. Esta métrica é importante para reduzir o número de **falsos positivos (FP)**, o que poderia levar a testes de acompanhamento desnecessários e ansiedade do paciente.
- **F1-score:** Definido como a média harmônica de *Precision* e *Recall*. Ele fornece uma medida única e **equilibrada** do desempenho de um modelo, o que é

particularmente útil em casos de desequilíbrio de classes.

## 4.2 Justificativa para Priorizar o *Recall*

Em um cenário de diagnóstico para uma condição crônica como a diabetes, um **falso negativo** (falhar em identificar uma pessoa com a doença) tem consequências para a saúde a longo prazo significativamente mais graves do que um **falso positivo** (sinalizar incorretamente uma pessoa saudável para testes adicionais, não invasivos). Um falso positivo leva a um reteste, enquanto um falso negativo pode levar a uma condição não tratada.

Por esta razão, o **Recall** foi designado como a **métrica primária** para determinar o *pipeline* superior. O modelo que demonstra a maior capacidade de identificar corretamente os pacientes com diabetes é considerado o mais valioso clinicamente.

A seção a seguir apresenta os resultados quantitativos da aplicação desta estrutura de avaliação a ambos os *pipelines*.

---

## 5.0 Resultados e Análise Comparativa

Esta seção apresenta os resultados quantitativos do experimento. As métricas de desempenho para o Pipeline A e o Pipeline B são resumidas, seguidas por uma comparação direta para identificar a abordagem mais eficaz com base na estrutura de avaliação estabelecida na seção anterior.

### 5.1 Resumo do Desempenho Quantitativo

Os principais resultados de desempenho para ambos os *pipelines* são resumidos na tabela abaixo. Os valores representam o intervalo de desempenho aproximado observado durante a execução do modelo.

Métrica de Desempenho	Pipeline A (Linha de Base)	Pipeline B (Otimizado)
Melhor Modelo (por <i>Recall</i> )	KNeighbors	Random Forest

<b>Máximo <i>Recall</i> (Melhor)</b>	$\approx 0.82 - 0.87$	$\approx 0.89 - 0.91$
<b>Melhor <i>F1-score</i></b>	$\approx 0.65 - 0.70$	$\approx 0.70 - 0.72$

## 5.2 Análise do Desempenho Comparativo

Os resultados empíricos demonstram uma **vantagem de desempenho decisiva para o Pipeline B**.

- O **Pipeline B** demonstra um desempenho superior na métrica primária, **Recall**, atingindo uma pontuação máxima no intervalo de **0.89-0.91**. Esta é uma melhoria notável em relação ao melhor desempenho do Pipeline A de  $0.82-0.87$ , significando uma maior capacidade de identificar corretamente pacientes com diabetes.
- Esse aumento na sensibilidade foi alcançado com um **impacto gerenciável na *precision***, como evidenciado pelo *F1-score* estável. Isso demonstra uma **troca bem-sucedida e deliberada**, priorizando a captura de casos verdadeiros positivos enquanto controla a taxa de falsos alarmes.
- O **Pipeline B** também atinge um *F1-score* ligeiramente superior, indicando que seus ganhos no *recall* foram alcançados mantendo um forte equilíbrio geral.
- O modelo com melhor desempenho para atingir o *Recall* máximo diferiu entre os dois *pipelines*: **KNeighbors** foi o de melhor desempenho no Pipeline A, enquanto **Random Forest** produziu os melhores resultados no Pipeline B.

Estes resultados numéricos destacam um vencedor claro. A seção a seguir passa destas descobertas quantitativas para uma discussão mais aprofundada sobre por que esses resultados ocorreram e suas implicações.

---

## 6.0 Discussão e Interpretação

Embora os resultados quantitativos apontem para um vencedor claro, é crucial entender as razões subjacentes para a diferença de desempenho. Esta seção interpreta os resultados conectando-os às **decisões arquitetônicas específicas** tomadas no Pipeline B e discute as implicações mais amplas de equilibrar a complexidade do modelo com os ganhos de desempenho em uma aplicação clínica.

## 6.1 Impacto da Otimização no *Recall*

O *Recall* superior do **Pipeline B** é um resultado direto de suas escolhas de *design* específicas. A **engenharia de recursos** (*feature engineering*) permitiu que o modelo aprendesse relacionamentos não lineares mais complexos indicativos da doença, criando variáveis de interação como *idade\_imc*. Concomitantemente, o uso de um **limiar de classificação mais baixo (0.3)** ajustou deliberadamente o limite de decisão do modelo para se alinhar com a prioridade clínica de minimizar os casos perdidos. Essa combinação foi o principal fator para o aumento de sua sensibilidade na captura de casos verdadeiros positivos.

## 6.2 Interpretabilidade do Modelo e Importância dos Recursos

Para garantir a **confiança clínica** e a **validade do modelo**, priorizamos a interpretabilidade do Pipeline B mais complexo. Técnicas como **Importância dos Recursos** (*Feature Importance*) e **gráficos SHAP** (*SHAP plots*) foram empregadas para validar a lógica do modelo. Essas análises confirmaram que os recursos de interação projetados (por exemplo, *idade\_imc*, *glicose\_imc*) possuíam alto valor preditivo, fornecendo evidências empíricas de que a etapa de engenharia de recursos foi eficaz e clinicamente relevante.

## 6.3 Análise de Complexidade *versus* Ganho de Desempenho

Uma questão central para este projeto era se o ganho de desempenho oferecido pelo Pipeline B justifica sua complexidade adicional. O **custo marginal** do aumento da complexidade de implementação — principalmente o tempo do desenvolvedor para engenharia e calibração de recursos — é **esmagadoramente compensado** pelo **valor clínico** da redução de falsos negativos. Em um contexto de saúde, o custo de um diagnóstico perdido não é medido em recursos computacionais, mas em potenciais resultados adversos para o paciente, tornando o *recall* superior do Pipeline B uma **vantagem inegociável**.

Esta interpretação fornece a justificativa para a recomendação final apresentada na seção de conclusão.

---

## 7.0 Conclusão e Recomendação

Este relatório detalhou uma análise comparativa de dois *pipelines* de *machine learning* para diagnóstico de diabetes. A análise conclui que uma estratégia de otimização direcionada produz uma melhoria **substancial e clinicamente significativa** na métrica de desempenho mais relevante.

### 7.1 Resumo das Descobertas

Os principais resultados da análise comparativa são sintetizados abaixo:

- O **Pipeline Otimizado (B)** superou consistente e significativamente o **Pipeline Linha de Base (A)** na métrica clínica mais crítica, **Recall**.
- Essa melhoria de desempenho foi diretamente atribuível a técnicas de otimização específicas, a saber, **engenharia de recursos avançada** e um **ajuste estratégico do limiar de classificação** para priorizar a sensibilidade.
- O Pipeline B também demonstrou um *F1-score* ligeiramente superior, indicando que seus ganhos no *Recall* foram alcançados mantendo um equilíbrio robusto geral entre precisão e sensibilidade.

### 7.2 Recomendação Final

Com base nesta análise abrangente, a recomendação inequívoca é **adotar o Pipeline B como a arquitetura fundamental** para o módulo de diagnóstico de diabetes do hospital. O trabalho futuro deve se concentrar na validação deste *pipeline* em relação aos dados internos dos pacientes e na sua preparação para **ensaios clínicos**.

Embora os resultados sejam promissores, é importante reconhecer o contexto e as restrições deste estudo.

---

## 8.0 Limitações do Projeto

Embora os resultados deste projeto sejam altamente encorajadores, é importante reconhecer as limitações do estudo atual para orientar futuros esforços de desenvolvimento e validação.

- **Fonte de Dados:** O modelo foi treinado exclusivamente no conjunto de dados



público *Pima Indians Diabetes Dataset*. Este conjunto de dados pode não ser totalmente representativo da população de pacientes específica do hospital, o que pode afetar o desempenho no mundo real.

- **Desequilíbrio de Classes Inerente:** O conjunto de dados tem um desequilíbrio natural entre casos positivos e negativos. Embora mitigado usando a técnica **SMOTE** em ambos os *pipelines*, isso continua sendo uma consideração crítica para a calibração do modelo e o desempenho no mundo real.
  - **Validação Clínica:** Os modelos ainda não foram submetidos a **validação clínica direta**. As percepções e previsões devem ser revisadas e verificadas por profissionais médicos qualificados antes de qualquer consideração para uso clínico.
  - **Escopo dos Recursos:** O conjunto de dados usado para este projeto carece de variáveis socioeconômicas potencialmente relevantes ou contextuais mais amplas que poderiam aprimorar ainda mais a precisão diagnóstica.
- 

## 9.0 Apêndices

### 9.1 Contribuintes do Projeto

Este projeto foi um esforço colaborativo dos seguintes membros da equipe:

- Adriel Santos
- João Marcos
- Leticia Nepomucena
- Lucas Silva
- Vagner Barbosa

### 9.3 Arquivos gerados pelo projeto

Todo o material utilizado no projeto se encontra no seguinte endereço do Github:

<https://github.com/vagnerbarbosa/tech-challenge-fase-1>

O Vídeo demonstrativo <https://www.youtube.com/watch?v=d62wPNKj1TU>

### 9.2 Ética e Conformidade de Dados

Este projeto foi conduzido em **aderência aos padrões de ética de dados**, utilizando exclusivamente um conjunto de dados público e anonimizado para todo o treinamento e

análise.