# **Relatório Técnico: Análise Preditiva de Diabetes**

**Tech Challenge - Fase 1**

**Autores:**

Ackeley Lennon (RM366072)

Antonio Rafael Ortega (RM365237)

Eduardo Tadeu Agarbella (RM366322)

Leandro Pessoa de Souza (RM365755)

Mateus Teixeira Castro (RM366469)

**Curso:** FIAP - IA para DEVS - 6IADT

**Data:** 12 de Agosto de 2025

## **1. Introdução**

### **1.1. Contexto e Objetivos**

O presente relatório detalha o desenvolvimento de um modelo de Machine Learning para a predição de diabetes, como parte do Tech Challenge da Pós-Graduação. O principal objetivo é criar uma solução de suporte ao diagnóstico que possa auxiliar equipes médicas na triagem e identificação de pacientes com alto risco de desenvolver a doença.

O diagnóstico precoce do diabetes é crucial para a prevenção de complicações graves e para a melhoria da qualidade de vida dos pacientes. Nesse contexto, um sistema inteligente de análise de dados pode otimizar o tempo dos profissionais de saúde e direcionar recursos para os casos que mais necessitam de atenção.

### **1.2. Metodologia**

O projeto seguiu uma metodologia estruturada de ciência de dados, compreendendo as seguintes etapas:

1. **Análise e Limpeza dos Dados:** Investigação da qualidade do dataset, com foco no tratamento de valores inconsistentes.
2. **Análise Exploratória (EDA):** Utilização de técnicas de visualização para extrair insights e entender as relações entre as variáveis.
3. **Pré-processamento e Modelagem:** Preparação dos dados para os algoritmos e treinamento de modelos de classificação, com especial atenção ao desbalanceamento de classes.
4. **Avaliação e Interpretação:** Análise crítica do desempenho dos modelos com métricas apropriadas para o contexto médico.

## **2. Descrição do Dataset**

O estudo foi realizado com base no dataset **Pima Indians Diabetes Database**, obtido através da plataforma Kaggle.

* **Fonte:** [Kaggle](https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data) - Diabetes Dataset
* **Descrição:** O conjunto de dados é composto por 768 registros de pacientes do sexo feminino, com 21 anos ou mais, de ascendência indígena Pima.

### **2.1. Dicionário de Dados**

A tabela a seguir descreve cada uma das variáveis presentes no dataset:

| **Coluna** | **Descrição em Português** |
| --- | --- |
| **Pregnancies** | Número de vezes que a paciente esteve grávida. |
| **Glucose** | Concentração de glicose no plasma 2 horas após um teste oral de tolerância à glicose. |
| **BloodPressure** | Pressão arterial diastólica (mm Hg). |
| **SkinThickness** | Espessura da dobra cutânea do tríceps (mm). |
| **Insulin** | Nível de insulina no soro após 2 horas (mu U/ml). |
| **BMI** | Índice de Massa Corporal (calculado como peso em kg / (altura em m)²). |
| **DiabetesPedigreeFunction** | Função de pedigree de diabetes, que representa uma pontuação da probabilidade de diabetes com base no histórico familiar. |
| **Age** | Idade da paciente (em anos). |
| **Outcome** | Variável alvo que indica o diagnóstico (0 = Não Diabético, 1 = Diabético). |

## **3. Análise Exploratória e Visualização de Dados**

### **3.1. Tratamento de Dados Inconsistentes**

Uma análise inicial revelou que diversas colunas (Glucose, BloodPressure, BMI, etc.) continham o valor 0, o que é fisiologicamente impossível. Estes valores foram tratados como dados ausentes e substituídos pela **mediana** de suas respectivas colunas, uma abordagem robusta que preserva a distribuição original dos dados.

### **3.2. Análise da Variável Alvo**

O dataset apresentou um desbalanceamento de classes, com aproximadamente 65% dos registros sendo de pacientes não diabéticos (Outcome=0) e 35% de pacientes diabéticos (Outcome=1). Essa característica foi fundamental para a escolha das estratégias de modelagem.

### **3.3. Análise de Fatores de Risco**

A análise de cruzamento de dados revelou insights importantes:

* **Glicose e IMC:** Pacientes com diabetes estão predominantemente concentrados na área de alta glicose e alto IMC, confirmando a forte interação entre esses dois fatores de risco.
* **Idade e Pressão Arterial:** A pressão arterial tende a aumentar com a idade para ambos os grupos. No entanto, a média de pressão para o grupo diabético é consistentemente maior.
* **Genética:** A função de pedigree (DiabetesPedigreeFunction) mostrou-se, em média, mais elevada para pacientes diabéticos, independentemente do número de gestações, reforçando o componente genético da doença.

## **4. Pré-processamento e Modelagem**

### **4.1. Estratégias de Pré-processamento**

1. **Divisão dos Dados:** O dataset foi dividido em 80% para treino e 20% para teste, utilizando a estratificação para manter a proporção das classes em ambos os conjuntos.
2. **Padronização (Standardization):** Foi aplicado o StandardScaler para normalizar as features, colocando todas na mesma escala (média 0 e desvio padrão 1). Isso é essencial para algoritmos sensíveis à escala, como a Regressão Logística.

### **4.2. Modelos Utilizados e Justificativa**

Foram treinados dois modelos de classificação: **Regressão Logística** e **Árvore de Decisão**.

A escolha principal da estratégia foi o uso do parâmetro class\_weight='balanced' em ambos os modelos. Esta técnica ajusta os pesos das classes, penalizando mais os erros cometidos na classe minoritária (pacientes diabéticos). O objetivo foi otimizar o modelo para a **detecção de casos positivos (alto recall)**, a métrica mais importante para um sistema de triagem médica.

## **5. Resultados e Avaliação**

A avaliação dos modelos foi realizada no conjunto de teste. A seguir, apresentamos os resultados do modelo com melhor desempenho: a **Regressão Logística Balanceada**.

| **Métrica** | **Não Diabético** | **Diabético** | **Acurácia Geral** |
| --- | --- | --- | --- |
| **Precision** | 0.82 | 0.60 | **73%** |
| **Recall** | 0.75 | **0.70** |  |
| **F1-Score** | 0.79 | 0.65 |  |

### **5.1. Interpretação dos Resultados**

* **Recall (70%):** O resultado mais importante é o recall de 70% para a classe "Diabético". Isso significa que o modelo foi capaz de identificar corretamente 7 em cada 10 pacientes que realmente tinham a doença. Para uma ferramenta de triagem, este é um indicador de alta eficácia.
* **Precision (60%):** A precisão de 60% para a classe "Diabético" indica que, quando o modelo prevê um caso como positivo, ele está correto em 60% das vezes. Os 40% restantes são falsos positivos, um trade-off aceitável para garantir que menos casos reais sejam perdidos.
* **Acurácia (73%):** A acurácia geral do modelo é de 73%, um valor sólido que demonstra a capacidade geral de classificação do modelo.

## **6. Conclusão**

O modelo de **Regressão Logística com Balanceamento de Classes** provou ser uma solução robusta e eficaz para a predição de diabetes neste contexto. Ao priorizar a métrica de recall, garantimos que o modelo sirva ao seu propósito principal: minimizar o número de pacientes doentes não identificados.

O modelo pode ser utilizado na prática?

Sim. Ele tem grande potencial para ser implementado como uma ferramenta de suporte à decisão em ambientes clínicos. O sistema pode analisar automaticamente os dados dos pacientes e gerar um alerta de risco, permitindo que a equipe médica priorize os casos que necessitam de investigação aprofundada. É fundamental ressaltar que o modelo não substitui o diagnóstico médico, mas o apoia, tornando o processo de triagem mais eficiente.

### **6.1. Próximos Passos e Melhorias**

* **Otimização de Hiperparâmetros:** Utilizar técnicas como GridSearchCV para refinar os parâmetros do modelo.
* **Modelos Avançados:** Testar algoritmos mais complexos, como RandomForest ou XGBoost, que podem capturar relações não-lineares nos dados.
* **Validação Cruzada:** Implementar uma validação cruzada mais robusta para garantir a generalização do modelo.

## **7. Apêndice**

O código-fonte completo, o Dockerfile e as instruções de execução estão disponíveis no seguinte repositório Git:

* **Link do Repositório:** <https://github.com/antrafa/fiap-tech-challenge-6IADT/>