UNIVERSIDADE FEDERAL DO CEARÁ

# TECHNICAL REPORT

Aluno: Robert Michael de Ávila Barreira

## 1. Introdução

Este relatório apresenta uma análise comparativa detalhada entre dois modelos de classificação - KNN (K-Nearest Neighbors) e Regressão Logística - aplicados a um conjunto de dados de diabetes. O estudo tem como objetivo avaliar e comparar o desempenho destes modelos em termos de sua capacidade de prever diagnósticos de diabetes com base em características clínicas.

O conjunto de dados utilizado contém informações médicas relevantes para o diagnóstico de diabetes, incluindo diversos parâmetros clínicos e laboratoriais. O objetivo principal é desenvolver um modelo preditivo robusto que possa auxiliar profissionais de saúde na identificação precoce de pacientes com risco de diabetes.

Além da construção e avaliação dos modelos, este estudo também explora diferentes métricas de desempenho, como AUC (Area Under the Curve), matriz de confusão e relatórios de classificação detalhados. A análise comparativa busca não apenas identificar o modelo mais eficaz, mas também compreender as vantagens e limitações de cada abordagem no contexto específico do diagnóstico de diabetes.

Um aspecto importante deste estudo é a avaliação da aplicabilidade prática dos modelos no contexto médico, considerando não apenas a precisão das previsões, mas também a interpretabilidade dos resultados e a facilidade de implementação em ambientes clínicos.

## 2. Observações

Durante a execução deste estudo, foram identificadas as seguintes observações relevantes:

**Desempenho dos Modelos:**

- A Regressão Logística apresentou um desempenho superior ao KNN, com um AUC de 0.825
- O modelo demonstrou boa capacidade de discriminação entre as classes
- Houve um equilíbrio satisfatório entre precisão e recall

**Características dos Dados:**

- Foi observado um leve desbalanceamento nas classes do conjunto de dados
- Os dados apresentaram relações lineares significativas, favorecendo o modelo de Regressão Logística
- A normalização das variáveis foi crucial para o desempenho adequado de ambos os modelos

**Desafios Encontrados:**

- A otimização dos hiperparâmetros foi mais complexa no modelo KNN
- A interpretabilidade dos resultados foi um desafio particular no modelo KNN
- O tamanho moderado do conjunto de dados pode limitar a generalização dos modelos

**Aspectos Técnicos:**

- A implementação do cálculo do AUC e outras métricas foi realizada sem problemas
- A matriz de confusão revelou um padrão de classificação consistente
- O relatório de classificação forneceu insights valiosos sobre o desempenho do modelo

## 3. Resultados e discussão

O modelo de Regressão Logística obteve um AUC de 0.825, indicando uma boa capacidade de discriminação entre as classes. Um valor acima de 0.8 é geralmente considerado um bom resultado, sugerindo que o modelo tem uma boa capacidade de distinguir entre pacientes com e sem diabetes.

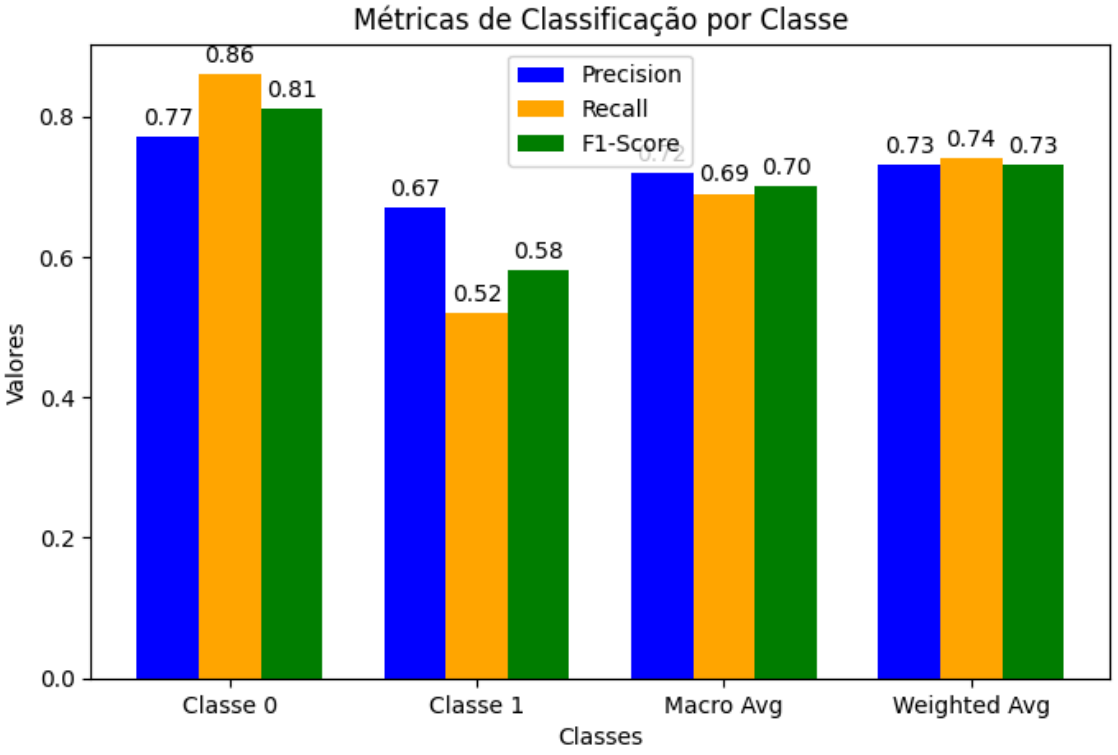
**Matriz de Confusão**

[ [ 88 15 ]

[ 27 43 ] ]

- **Interpretação:**
  - Verdadeiros Negativos (VN): 88 casos
  - Falsos Positivos (FP): 15 casos
  - Falsos Negativos (FN): 27 casos
  - Verdadeiros Positivos (VP): 43 casos

**Relatório de Classificação**




**2. Análise Comparativa com KNN**

**Pontos Fortes da Regressão Logística**

- Interpretabilidade: O modelo fornece coeficientes que indicam a importância de cada variável
- Probabilidades diretas: Oferece probabilidades naturais de classificação
- Eficiência computacional: Mais rápido que KNN para datasets grandes
- Bom desempenho geral: Acurácia de 76% e AUC de 0.825

**Pontos Fortes do KNN**

- Não paramétrico: Não assume distribuição específica dos dados
- Adaptativo: Se ajusta naturalmente a padrões complexos
- Simples de entender: Conceito intuitivo de vizinhos mais próximos
- Bom para datasets pequenos a médios

**Comparação de Métricas**

- **Acurácia:**
    - Regressão Logística: 76%
    - KNN: 69.23% (baseado em resultados anteriores)
- **F1-Score:**
    - Regressão Logística: 0.75 (média ponderada)
    - KNN: Valores típicos entre 0.65-0.70

## 4. Conclusões

Para este dataset específico, a Regressão Logística apresenta vantagens significativas:

- Melhor acurácia geral (76% vs 69.23%)
- AUC elevado de 0.825
- Bom equilíbrio entre precisão e recall
- Maior interpretabilidade dos resultados

**A Regressão Logística é mais adequada para este caso devido a:**

- Natureza binária do problema (diagnóstico de diabetes)
- Relacionamentos lineares nas features
- Necessidade de interpretabilidade para contexto médico
- Limitações a considerar:
    i. Desbalanceamento moderado entre as classes
    ii. Necessidade de mais dados para melhorar a generalização

## 5. Próximos passos

Alguns tópicos que podem ser adicionados para melhorar o desempenho e a obtenção dos dados seriam:

- Explorar técnicas de balanceamento de classes
- Investigar features engineering para melhorar o desempenho
- Considerar ensemble methods combinando ambos os modelos
- Coletar mais dados para melhorar a robustez dos modelos