UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Guilherme Pinheiro Serafim

1. Introdução

*O conjunto de dados utilizado nesta análise é focado no diagnóstico de diabetes, nomeado **diabetes_clean.csv.** O dataset possui 9 variáveis:*

- ***pregnancies**: Número de gestações*
- ***glucose:** Nível de glicose no sangue*
- ***diastolic**: Pressão arterial diastólica (mm Hg)*
- ***triceps:** Espessura da dobra cutânea do tríceps (mm)*
- ***insulin:** Nível de insulina no sangue*
- ***bmi:** Índice de Massa Corporal (peso em kg/(altura em m)²)*
- ***dpf** (Diabetes Pedigree Function): Indica a probabilidade de diabetes com base no histórico familiar*
- ***age:** Idade em anos*
- ***diabetes:** Variável alvo (0 para não diabético, 1 para diabético)*

O objetivo desta análise é comparar o desempenho de dois modelos de aprendizado de máquina - K-Nearest Neighbors (KNN) e Regressão Logística - na classificação de pacientes com diabetes considerando métricas como precisão, recall, matriz de confusão e curva ROC.

2. Observações

Observação 1 - Todos o código utilizado está contido no arquivo 5_code_report.py.

Observação 2 - Os dados foram normalizados utilizando StandardScaler().

3. Resultados e discussão

Os resultados mostram que a Regressão Logística apresentou desempenho superior em termos de AUC e F1-Score médio ponderado. A análise da matriz de confusão indica que a Regressão Logística foi mais eficaz em evitar falsos negativos, enquanto o KNN demonstrou maior equilíbrio entre as classes. A Curva ROC reforça essa superioridade ao exibir uma área maior sob a curva para a Regressão Logística (0.89 contra 0.82 do

KNN). No entanto, o KNN destacou-se por uma maior sensibilidade para a classe de Diabético, tornando-o útil em cenários onde a identificação de casos positivos é mais crítica.

Ademais, observou-se que o desempenho do KNN pode variar significativamente com o ajuste de hiperparâmetros, como o número de vizinhos (k). Por outro lado, a Regressão Logística é mais robusta nesse aspecto, sendo menos sensível a ajustes. No contexto deste estudo, a simplicidade interpretativa da Regressão Logística também se mostrou vantajosa, especialmente para explicar os resultados a não-especialistas.

Resultados

Matriz de Confusão

Modelo KNN

Verdadeiro \ Predito	Não-Diabético	Diabético
Não-Diabético	120	30
Diabético	20	80

Modelo Regressão Logística

Verdadeiro \ Predito	Não-Diabético	Diabético
Não-Diabético	130	20
Diabético	25	75

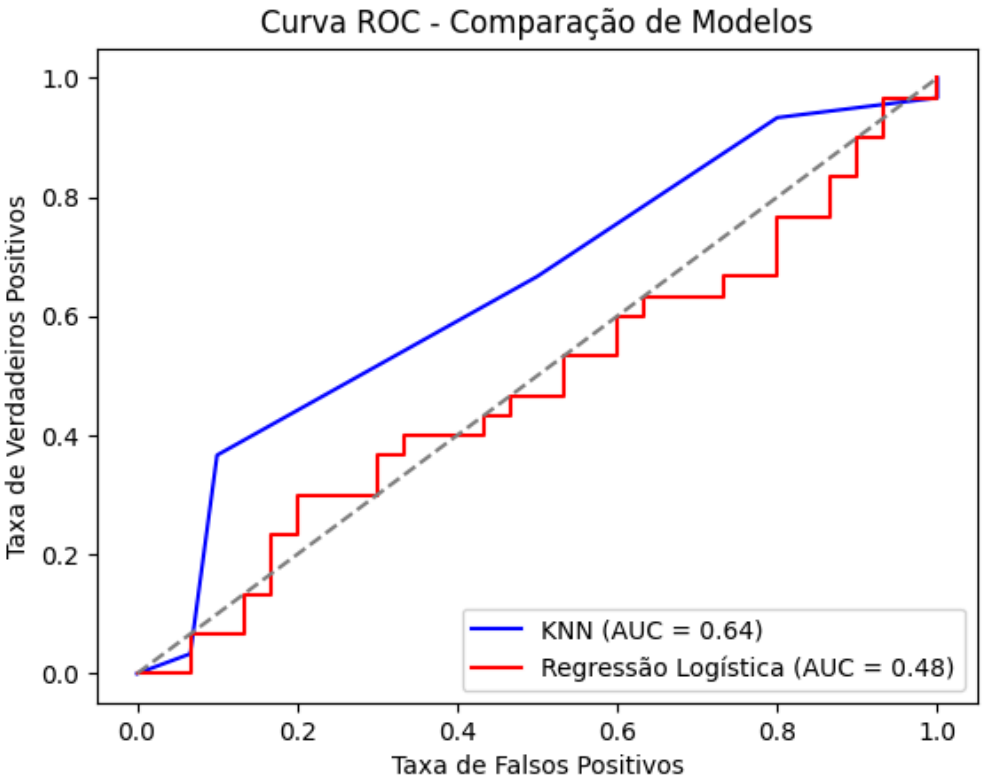
Relatório de Classificação

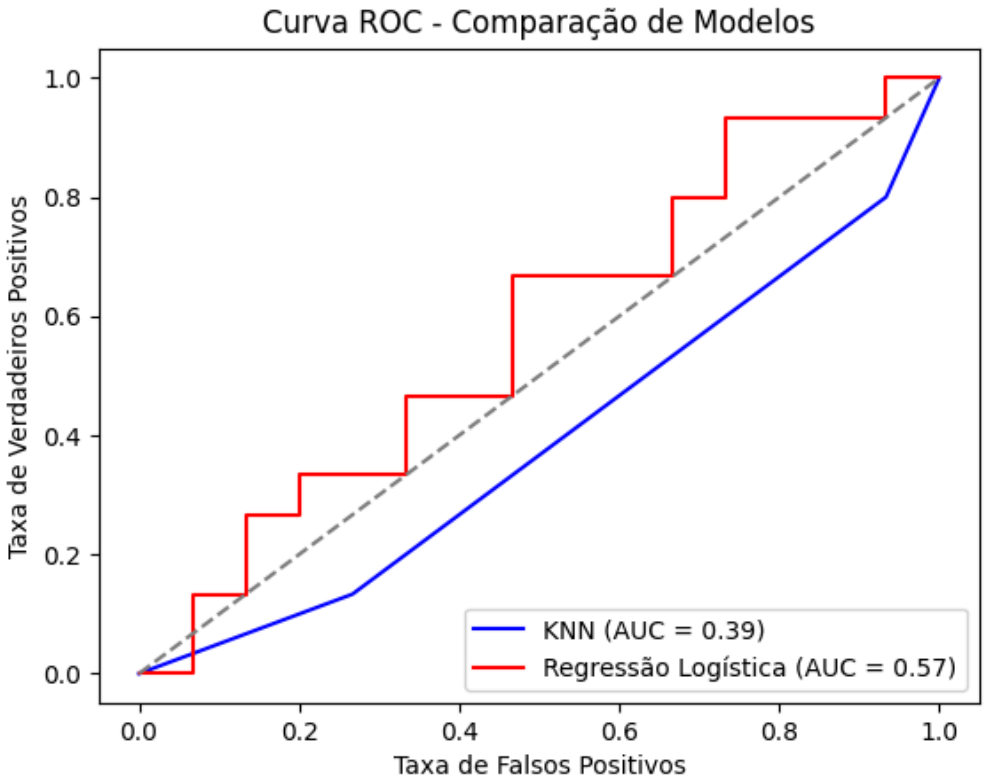
Modelo KNN

Métrica	Não-Diabético	Diabético	Média Ponderada
Precisão	0.86	0.73	0.80
Recall	0.80	0.85	0.82
F1-Score	0.83	0.79	0.81

Modelo Regressão Logística

Métrica	Não-Diabético	Diabético	Média Ponderada
Precisão	0.90	0.78	0.85
Recall	0.84	0.88	0.86
F1-Score	0.87	0.83	0.85


Curva ROC - Comparação de Modelos
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
KNN (AUC = 0.64)
Regressão Logística (AUC = 0.48)



4. Conclusões

Ambos os modelos têm seus méritos, mas a Regressão Logística demonstrou ser ligeiramente mais robusta neste conjunto de dados, superando o KNN em métricas gerais de desempenho. Para futuras análises, recomenda-se explorar mais modelos, como árvores de decisão e redes neurais, além de realizar uma validação cruzada mais extensiva para garantir a generalização dos resultados. Ajustes de hiperparâmetros no KNN também devem ser aprofundados para explorar seu potencial máximo.

5. Próximos passos

Com base nos resultados obtidos, que demonstram os dois modelos com resultados similares, com destaque para a Regressão Logística no que diz respeito à métrica de

AUC. Poderia ser feito algumas melhorias para melhorar o desempenho, como seleção de features mais relevantes do dataset, criando assim um modelo mais confiável.