UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luan Sales Barroso

1. Introdução

O dataset se chama *diabets_clean.csv* ele contém alguns dados de saúde como por exemplo: glucose, insulin, bmi, age, entre outros. Que tem relação com a diabete, por meio desses dados é definido se a pessoa tem ou não diabetes.

2. Observações

3. Resultados e discussão

Para a resolução dessa questão bibliotecas essenciais como *KNeighborsClassifier(com 6 vizinhos), LogisticRegression, train_test_split* foram importadas para a realização da atividade, inicialmente o dataset foi iniciado e a coluna alvo escolhida foi 'diabetes', então a função *train_test_split* foi executada e armazenada nas seguintes variaveis: X_train, X_test, y_train, y_test. Então os dados de X train e test foram normalizados usando o StandardScaler para ter resultados mais controlados, em seguida o fit e a predição do knn e da regressão logística foram realizados, com isso o desempenho do knn e da regressão logística foram as seguintes:

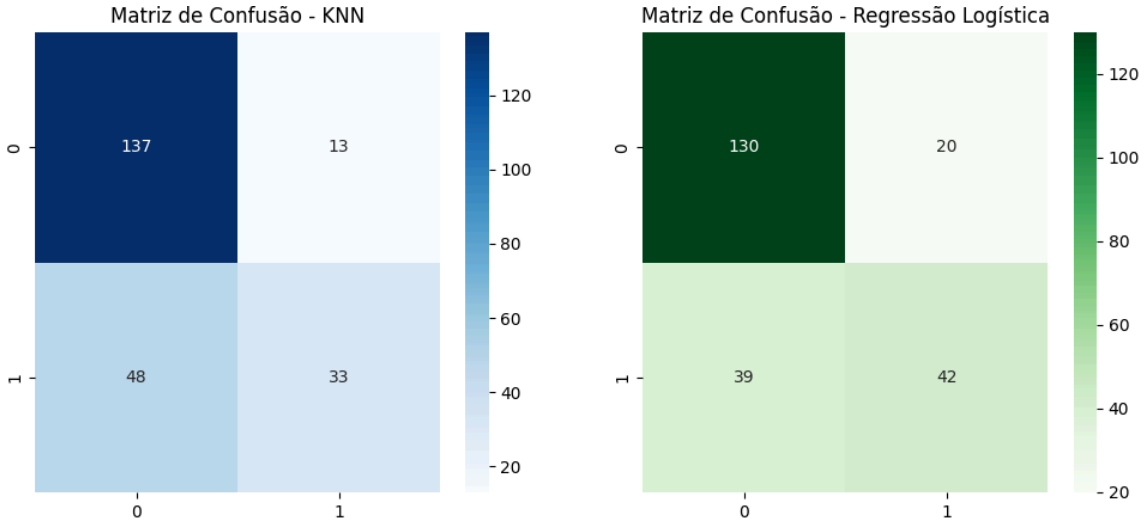
Desempenho do KNN:

	precision	recall	f1-score	support
0	0.74	0.91	0.82	150
1	0.72	0.41	0.52	81
accuracy			0.74	231
macro avg	0.73	0.66	0.67	231
weighted avg	0.73	0.74	0.71	231

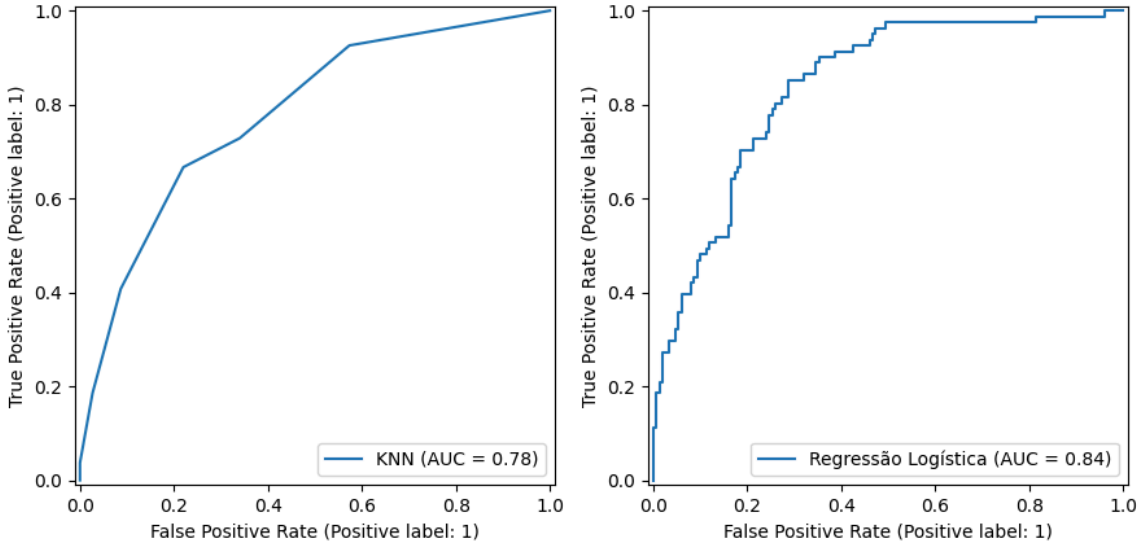
Desempenho da Regressão Logística:

	precision	recall	f1-score	support
0	0.77	0.87	0.82	150
1	0.68	0.52	0.59	81
accuracy			0.74	231
macro avg	0.72	0.69	0.70	231
weighted avg	0.74	0.74	0.74	231

Com base nesse primeiro resultado, a escolha depende bastante da sua necessidade, o KNN por exemplo apresenta uma precisão maior, então se seu foco for na precisão opte pelo KNN. Em seguida a matriz de confusão foi usada para demonstrar de forma mais visual o desempenho:



Com base nas matrizes, o KNN teve um desempenho melhor em relação aos verdadeiros positivos em comparação com a regressão logística, mas a regressão logística teve um desempenho melhor em relação aos verdadeiros negativos em comparação com o KNN. Por fim, foi gerado uma curva ROC para ambos os modelos:



Baseado no resultado da curva ROC, a Regressão logística teve um desempenho muito bom na distinção entre as classes mesmo tendo uma curva meio estranha e não tão suave. E como resultado final, a acurácia do KNN foi de 0.735931 enquanto a da Regressão logística foi de 0.744589 o que dá mais um ponto para a Regressão Logística.

4. Conclusões

Com base nos resultados apresentados, é possível concluir que até certo ponto, depende da sua necessidade (com base no dataset utilizado) se é precisão ou sensibilidade, mas em um âmbito geral a Regressão Logística apresentou um desempenho superior o que era esperado já que para o KNN ter um melhor desempenho é necessário a definição do melhor número de vizinhos para cada dataset.

5. Próximos passos

Imagino que seria interessante a adição de outros modelos a serem avaliados.