UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: José Eric Mesquita Coelho

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.

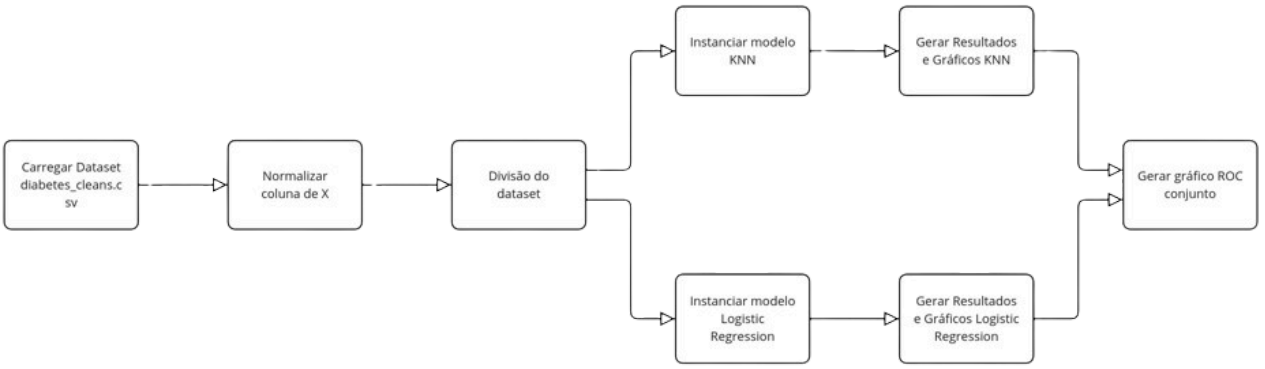
2. Observações

Observação 1 - Todos o código utilizado está contido no arquivo **5_code_report.py**, para melhor compreensão.

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

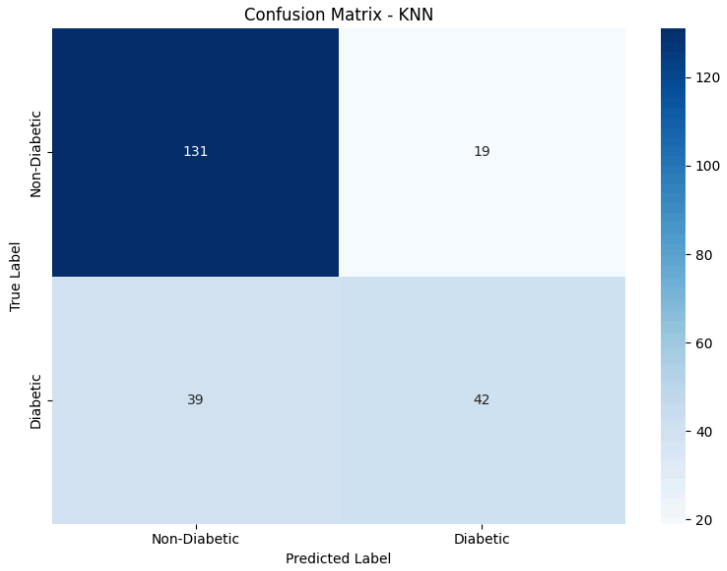
3. Resultados e discussão

Ambos os modelos foram avaliados usando um conjunto de testes contendo 231 amostras válidas (cerca de 30% do total do dataset). A execução dentro do código ocorre da seguinte maneira:

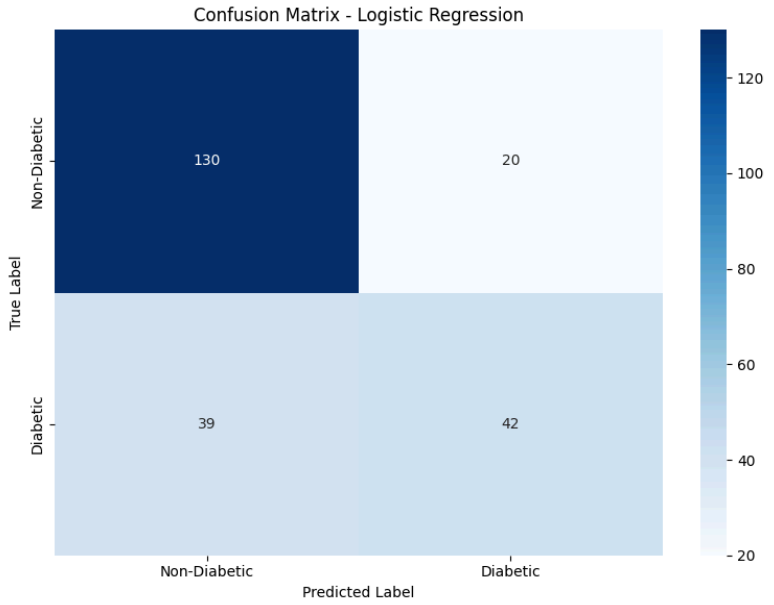


Análise da Matriz de Confusão:

1. KNN:



2. Regressão Logística:



Se observa uma similaridade notável entre os dois modelos, com apenas uma pequena diferença nos casos de verdadeiros negativos e falsos positivos. Isso possivelmente se deve ao fato da amostra de dados ser limitada.

Análise das Métricas de Desempenho:

1. KNN:

Métrica	Classe 0	Classe 1	Macro Avg	Weighted Avg	Overall
Precision	0.77	0.69	0.73	0.74	-
Recall	0.87	0.52	0.70	0.75	-
F1-Score	0.82	0.59	0.71	0.74	-
Accuracy	-	-	-	-	0.75
AUC	-	-	-	-	0.7856

2. Regressão Logística:

Métrica	Classe 0	Classe 1	Macro Avg	Weighted Avg	Overall
Precision	0.77	0.68	0.72	0.74	-
Recall	0.87	0.52	0.69	0.74	-
F1-Score	0.82	0.59	0.70	0.74	-
Accuracy	-	-	-	-	0.74
AUC	-	-	-	-	0.8380

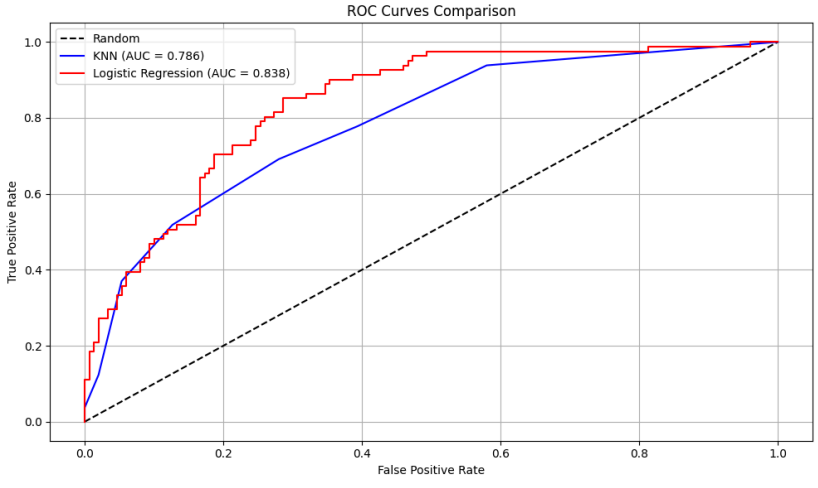
Ambos tiveram resultados bem semelhantes, possivelmente por conta da normalização e por o conjunto de dados estar bem adaptado aos dois modelos.

Análise da Curva ROC e AUC:

A diferença mais significativa entre os modelos aparece no score AUC:

- KNN: AUC = 0.7856
- Regressão Logística: AUC = 0.8380

A Regressão Logística apresentou um AUC superior, indicando melhor capacidade de discriminação entre as classes ao longo de diferentes limiares de classificação.



4. Conclusões

Em conclusão, embora os modelos apresentem desempenho similar em várias métricas, a Regressão Logística demonstrou vantagem no score AUC, sugerindo maior robustez na classificação geral. No entanto, ambos os modelos precisam de melhorias, especialmente na identificação de casos positivos, crucial em aplicações médicas.

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.