FATEC Baixada Santista - Rubens Lara

Tecnólogo em Ciência de Dados

Ellen Christine Ferreira Ozores

Relatório:
Predição de Diabetes com Regressão Logística e Gradiente Descendente

3º Ciclo - Cálculo

Santos

2025

INTRODUÇÃO

Esse relatório apresenta os resultados da aplicação da regressão logística em um dataset, utilizando o método de otimização gradiente descendente, a fim de prever o diagnóstico de diabetes entre mulheres com base em características clínicas das pacientes apresentadas no conjunto de dados. A regressão logística é um algoritmo estatístico ideal para prever probabilidades associadas a categorias, com isso, essa técnica utiliza uma função sigmoide para limitar a saída a valores entre 0 e 1, possibilitando a interpretação como uma probabilidade, que no caso deste trabalho, se resume em diabético e não diabético. Já o gradiente descendente, é uma técnica de otimização usada para encontrar os parâmetros que minimizam a função de custo do modelo, permitindo ajustar os coeficientes da regressão para que a previsão feita se aproxime ao máximo dos valores reais.

DESCRIÇÃO DO DATASET

O conjunto de dados utilizado neste trabalho é o *PIMA Indians Diabetes Database*, disponibilizado pela National Institute of Diabetes and Digestive and Kidney Diseases e disponível gratuitamente no Kaggle. O dataset contém informações de 768 pacientes, mulheres de origem Pima (uma população indígena dos Estados Unidos), apresentando variáveis sobre número de gestações anteriores, nível de glicose no sangue, pressão arterial diastólica, concentração sérica de insulina, índice de massa corporal, histórico familiar de diabetes e idade, que foram usados como atributos preditores e como variável alvo, foi usada "outcome", denominadas com 0 ou 1.

METODOLOGIA

O modelo de regressão logística foi implementado utilizando a linguagem Python e as bibliotecas NumPy, Pandas, Matplotlib e scikit-learn. O algoritmo consiste em encontrar os parâmetros que minimizam uma função de custo, ajustando os pesos por meio de iterações do gradiente descendente.

Para melhorar a performance do gradiente descendente, os dados foram normalizados (padronização das variáveis para valores entre 0 e 1) e o conjunto foi dividido entre dados de treino (70%) e dados de teste (30%), para garantir uma avaliação justa do modelo.

Além disso, algumas etapas e ferramentas também foram usadas no trabalho:

a) Função sigmoide: transforma a saída dos dados em uma probabilidade entre 0 e 1.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{com } z = X\theta$$

b) Função de custo: foi usada a "log-loss", que mede o erro entre as previsões e os valores reais.

c) Gradiente Descendente e atualização dos pesos.

Após o treinamento, o modelo foi analisado seguindo as métricas encontradas na acurácia, precisão, recall (sensibilidade) e o f1-score.

PROBLEMA DO MÍNIMO

Neste trabalho, o problema do mínimo foi minimizar a função de custo da regressão logística para que o modelo pudesse fazer previsões mais precisas sobre diabetes, através do gradiente descendente, até que o modelo convergisse para uma solução estável.

$$\min_{\theta} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \cdot \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \cdot \log(1 - \hat{y}^{(i)}) \right]$$

J(θ): função de custo que queremos minimizar

n: número de exemplos no conjunto de dados

$y^{(i)}$**:** valor real (0 ou 1) do exemplo i

$\hat{y}^{(i)}$**:** previsão feita pelo modelo para o exemplo i

RESULTADOS OBTIDOS

Após o treinamento do modelo, ao avaliar o desempenho do conjunto do teste, obtivemos os seguintes resultados:

a) **Acurácia:** 0,75

A acurácia parece alta ao ser analisada isoladamente, porém ela consegue ser influenciada pela classe majoritária do modelo, sendo assim, nesse caso específico, pode não ser a melhor métrica para analisar o desempenho.

b) **Recall:** 0,67

Quando se trata de predição de doenças, o Recall é uma métrica essencial para a análise, pois ele mede a proporção de verdadeiros positivos em relação ao todo, no caso do trabalho, se aplica em pacientes com diabetes que foram corretamente diagnosticados, que segundo o Recall, o acerto ficou em torno de 67%. Em um contexto médico, é fundamental detectar a maioria dos casos de diabetes para garantir o tratamento adequado, sendo assim, mesmo com o Recall razoável, é importante que o valor seja maior.

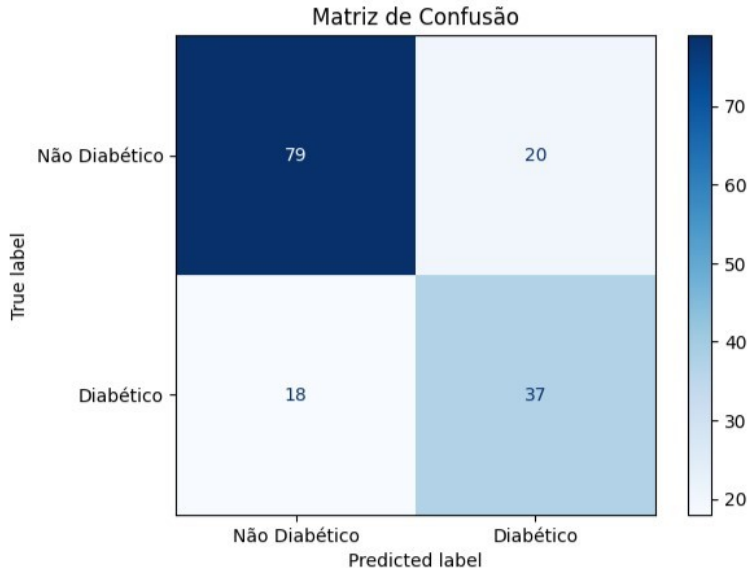
c) **Precisão:** 0,64

O modelo tem uma precisão de 64%, o que significa que cerca de 35% dos pacientes identificados como diabéticos, não têm a doença (falsos positivos). Esse fator pode ser um problema ao levar em consideração o tratamento desnecessário que muitos deles terão, justamente pelo diagnóstico errado.

d) **F1-Score:** 0,66

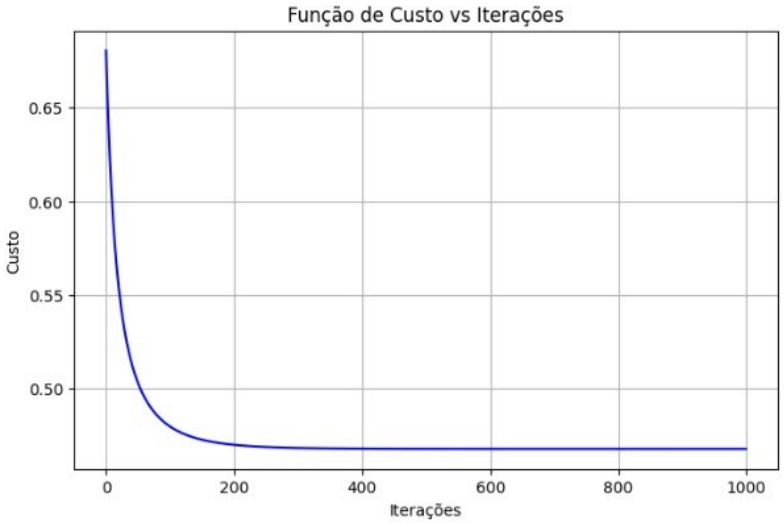
Essa métrica indica um bom desempenho geral do modelo, existindo um equilíbrio entre a precisão e o recall, no entanto, o valor ainda pode ser melhorado.

A figura 01 traz os dados encontrados na matriz de confusão.



Por meio dos números obtidos, conseguimos analisar que 79 dos pacientes foram previstos como não diabéticos e eles realmente não são, é o caso verdadeiros negativo, ou seja, o valor previsto pelo modelo foi equivalente com o real, além disso, realizou-se a previsão de 37 pacientes que têm diabetes e o real apontou que o diagnóstico é verdadeiro. Em relação aos falsos negativos, 18 pacientes foram colocados como não diabéticos sendo que eles possuem a doença e 20 são falsos positivos, isto é, receberam o diagnóstico mas não têm diabetes.

A figura 02 apresenta a função de custo em relação às interações do modelo.



A função de custo apresentou uma queda acentuada nas primeiras iterações, isso indica que o modelo conseguiu aprender rapidamente, atingindo uma convergência estável em

torno de 300. Com isso, há uma diminuição progressiva da função conforme o aumento das iterações, evidenciando a eficácia do gradiente descendente.

Além das métricas de desempenho, é possível avaliar a influência das variáveis no diagnóstico de diabetes. No caso da glicose, quanto maior a concentração, maior será a probabilidade de ter a enfermidade, sendo a variável mais determinante. O IMC (índice de massa corporal) também teve peso significativo, as mulheres com o maior número, apresentaram mais chances de diagnóstico positivo, e em relação à idade, houve uma tendência de aumento do risco conforme ela cresce.

CONCLUSÃO

O presente trabalho teve como objetivo aplicar a regressão logística com otimização via gradiente descendente para prever casos de diabetes em mulheres com base em dados clínicos, resultando em uma função de custo convergente e no desempenho equilibrado do modelo, mas não excelente. Ademais, o recall demonstrou ser ligeiramente maior que a precisão, indicando uma tendência maior a identificar pacientes com diabetes, porém, pode gerar mais falsos positivos também.

A análise entre a relação das variáveis como glicose, IMC e idade com a diabetes, revelou um forte impacto no diagnóstico, o que reforça a utilidade do modelo como ferramenta de apoio em triagens médicas. Ainda que simples, a regressão logística mostrou-se efetiva, interpretável e coerente com os padrões clínicos reais.

Dessa forma, esse trabalho evidencia que técnicas básicas de aprendizado de máquina, quando aplicadas, podem trazer resultados úteis e explicáveis, mesmo com implementações simples.

REFERÊNCIAS

- NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. *Pima Indians Diabetes Database*. Kaggle. Disponível em: `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`
- Link do GitHub, com o dataset utilizado e o código: `https://github.com/ellenozores/PredicaoDiabete_RegLog.git`